THE DEVELOPMENT AND VALIDATION OF A NOVEL TASK TO QUANTIFY FUNCTIONAL LANGUAGE PROFICIENCY IN SPANISH-ENGLISH LEARNING SCHOOL-AGE CHILDREN

By

Genesis D. Arizmendi

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF SPEECH, LANGUAGE, AND HEARING SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

UNIVERSITY OF ARIZONA

2019

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Genesis Arizmendi, titled 'The Development and Validation of a Novel Task to Quantify Functional Language Proficiency in Spanish-English Learning School-Age Children' and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Mary Alt, PhD, CCC-SLP — Date: 4-19-19

Elena Plante, PhD, CCC-SLP — Date: 4-19-19

Leah Fabiano-Smith, PhD, CCC-SLP — Date: 4-19-19

_____ — Date: _____

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Mary Alt, PhD, CCC-SLP — Date: 4-30-19
Dissertation Committee Chair
Speech, Language, and Hearing Sciences

ACKNOWLEDGEMENTS

This work would have not been possible without the encouragement and support from my mentors.

Dr. Alt, I am eternally grateful for your mentorship, guidance, and friendship over the last decade. Your encouragement from our my first rotation in the lab as an undergraduate inspired me to continue my studies and know that I was capable of achieving all that I've managed to throughout the years. Your support in both my career, as well as my personal growth, has led me to reach heights I never thought I could achieve. Thanks for taking a chance on a young border-town girl with no experience in research. I couldn't have dreamed for a better mentor.

Dr. Plante, your passion, dedication, and commitment to this profession has inspired me from day one. I am continuously striving to be all that I can be, hoping that one day I can inspire someone the way you have inspired me.

Dr. Fabiano-Smith, thank you for your support over the years and for being a strong model for how to be a prominent voice for underrepresented groups within our profession.

I am also grateful for one of the greatest friends I had the pleasure of making throughout my doctoral program, Trianna Oglivie. You are such a wonderful human being. I thank you for your friendship and support throughout this program. I am looking forward to all we can do together to improve the lives of children.

I also wish to express my greatest appreciation to my mother, Clarisa German, who has always shown and taught me the value of an education. Thank you for helping and supporting me through the good and bad times of my graduate program.

Lastly, I'd like to thank the funding sources that have made this degree possible. NIH/NIDCD, the Council of Academic Programs in Communication Sciences and Disorders, the University of Arizona National Institutes of Health Initiatives to Maximize Student Development Program, the University of Arizona Graduate College, The Grunewald Foundation, and the Louise Foucar Marshall Foundation.

DEDICATION

Para mi gente.

Que esto sea un ejemplo de hasta dónde tenemos la capacidad de llegar.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Clinicians, educators, and researchers alike continue to struggle without adequate and functional tools to measure language proficiency in bilingual populations. Language proficiency refers to the ability of an individual to use a language. However, the ways in which proficiency is classified are inconsistent and potentially invalid. Proficiency in young bilingual children is often determined through indirect measures (e.g., parent report) with unknown or inconsistent validity, impacting the field in both clinical and research arenas. The purpose of this study was to develop and validate a novel task that will allow us to quantify a child's functional language proficiency, while also identifying areas of language strengths and weakness across languages in Spanish-English 1st, 2nd, and 3rd grade children. The task capitalizes on the theory of natural translation (Harris & Sherwood, 1978), which refers to translation done in everyday circumstances by those who have had no special training. We evaluated task components and total task reliability and validity using test theory procedures. This work will set the foundation for quantifying and characterizing language proficiency in typically-developing Spanish-English speaking children.

Introduction

The problem of and need for quantifying language proficiency for culturally and linguistically diverse children

According to the U.S. Census Bureau population estimates as of 2017, there are roughly 58.9 million Hispanics living in the United States, constituting 18.1 percent of the nation's population. Of that population, there are over 37 million Hispanics who report that Spanish is the language of the home. Children who are born to parents who are Spanish speakers will inevitably become English language learners (ELLs), or in the process of becoming bilingual, upon entering the schools. As of fall 2015, there were 4.8 million children enrolled in public schools who were classified as being ELL, with 77% of those children speaking Spanish as their primary language (National Center for Education Statistics, 2018). The problem for these children is that they often display difficulties in school because of the extra cognitive load of learning a new language, all while learning the expectations, routine, and academic demands of being in a school environment (Miller & Endo, 2004).

These linguistic differences present a variety of challenges for children who are English language learners or for those who are bilingual. One of the primary reasons for poor academic performance is limited proficiency in English (Rumberger & Larson, 1998). Language proficiency refers to the ability of an individual to speak or perform in an acquired language. For example, if a child who began learning English upon entering school is tested in English three months into the school year, it is likely that the child will not be able to complete many tasks accurately due to poor English language comprehension, poor verbal production in English, or both. Indeed, Hoff and Core (2015) argue that "the data are clear that poor English skills at school entry place a child at risk for school failure" (p. 96). They struggle not because they do

not have the ability to master the academic material, but because they are only beginning the process of learning English and attempting to achieve English language proficiency in order to meet the demands of school. It is not until these children reach a certain level of English language proficiency that they can meet the academic demands being placed on them.

When English language learning children begin to fall behind academically, they may sometimes be referred for special education services and can be inaccurately diagnosed as needing these services based on performance on English-only measures. This leads to the reported overrepresentation of English language learning children in special education (e.g., Artiles, Rueda, Salazar, & Higareda, 2005; Sullivan, 2011). On the other hand, ELLs who have a legitimate need for special education services may not be referred simply because teachers suspect their academic struggles are only a language proficiency issue, leading to a disproportionate underrepresentation of minority children in special education (e.g., Morgan et. al, 2015; Morgan et. al, 2018). These reports lead to a problematic paradox. An ELL child's language development is often complex and poorly understood. This is magnified when educators, clinicians, and researchers have inadequate methods of testing and supporting the development of an ELL's linguistic abilities.

Though it is best practice to evaluate bilinguals in both of their languages (Bedore & Peña, 2008; Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004 [PL 108-446]); Kohnert, 2010), educators and clinicians continue to evaluate English skills only (Caesar & Kohler, 2007). This choice may be due to inadequate training in serving bilingual students (Hammer, Detwiler, Blood, & Qualls, 2004), lack of access to, or perceived lack of access to appropriate assessment materials or people with the ability to speak the child's non-English language (Roseberry-McKibbin, Brice, O'Hanlon, 2005), or to an overestimation of the

child's English language abilities. Assessing language skills only in English provides an incomplete snapshot of a child's overall linguistic development and abilities. Although some children may appear to grasp English fairly well early on, there are more complex language demands in the tests given to children, which affect overall performance. For example, ELL students generally perform lower than non-ELL students in reading, science, and math (Abedi, 2002). Indeed, "the level of impact of language proficiency on assessment of ELL students is greater in content areas with a higher level of language demand" (Abedi, 2002, p.232). This is particularly true for standardized testing, which often has high-stakes academic consequences, like retention of a grade or being placed in special education.

Standardized tests are often designed for the assessment of native monolingual English speakers, creating an even bigger problem in that children are essentially being tested on their language proficiency, rather than the content of the test (Menken, 2008). For example, in Alt, Arizmendi, Beal, and Hurtado (2013), ELL children were tested on the KeyMath-3, an English, standardized mathematics test. In English, children had significantly lower scores than their monolingual peers. However, when we adjusted for language, by administering a Spanish version of the test, the ELL children performed equally to their monolingual peers on these measures. This reflects the fact that ELL children did not have poorer understanding of mathematical concepts – the putative construct the test was designed to measure- but rather a poorer understanding of the language used in the assessment. Importantly, all ELL showed improved scores on the version of the test that allowed for Spanish, even those children who might have been classified as English dominant. In addition to not testing in both languages, there are also problems with the quality of some measures. Assessments that are typically administered to ELL children often are inappropriate to use for the population (e.g., poor overall

psychometric properties or norming populations that do not match the children tested). Because of inappropriate test selection, children's performance on assessments may not reflect their actual abilities (Wolf & Leon, 2009).

Considering that language proficiency plays such a strong role in the academic outcomes for children, it is surprising that there are not yet any agreed upon ways for quantifying proficiency. This is one of the, if not the most, fundamental questions to answer if we want to get to the root of helping this rapidly increasing population. However, when we think about what it means to be an ELL, or be in the process of becoming bilingual, the basic principles can become complicated.

Being bilingual means different things to different people. One person could define the term bilingual as being able to speak two languages fluently, while another could define it as the ability to speak or understand a bit of a second language. People do not typically ask, "What's your definition of bilingualism?" Assumptions are made about proficiency based on different, usually implicit, definitions. Bilingualism is a relative, rather than an absolute concept. Disparities in the definition of what it means to be bilingual translates to potential problems in classifying linguistic abilities and setting expectations for bilingual children in an educational setting. It also affects the way that we conduct science and interpret our findings with this population. Indeed, a range of official documents ranging from educational policies, published works, and literature reviews have consistently cited the lack of standard operational definitions for what it means to be a bilingual, or ELL (Ragan & Lesaux, 2006). A clear measure of proficiency could help not only define, but quantify the nature of a person's bilingualism at a given point in time.

The lack of reliable proficiency measures significantly impacts how research with bilingual populations is interpreted, given the variability in backgrounds and proficiency of the participants who are recruited for research (Grosjean, 1998). This is evident when we take a look at bilingual research. Let us take, for example, the "bilingual cognitive advantage" debate. The literature has documented cognitive advantages in bilinguals for decades (e.g., Bialystok & Martin, 2004; Diaz, 1985; Peal & Lambert, 1962). Recently, however, there has been both an increase in the number of studies that support these claims (e.g., Adesope, Lavin, Thompson, & Ungerleider, 2010) and those that refute them (e.g., Arizmendi et al., 2018; deBruin, Treccani, & Della Sala, 2015). Could one underlying reason for this discrepancy be that the types of bilinguals recruited are all defined differently with varying proficiency levels, leading to differences in the way language use affects the cognitive system?

There is no agreed upon and systematic measure among researchers who are conducting this work on what we mean by "bilingual." Some researchers use age of acquisition measures, others might select participants by input/output calculations, others use school-based classification of ELL, and yet others simply rely on the parent saying that the child is bilingual. The list could go on and on (e.g., Bedore et al.,2012). So why is this problem not being more directly addressed as a fundamental issue in clinical, educational, and research arenas? Considering that this population has been reported as being the fastest growing segment of the U.S. school-age population for over a decade (Genesee, Lindholm-Leary, Saunders, & Christian, 2005), it is time that the issue be seriously addressed.

Current approaches to quantify proficiency in Communication Sciences and Disorders

Clinicians, educators, and researchers struggle to measure language proficiency in bilingual populations due to a lack of adequate and functional tools. Without adequate measures to quantify linguistic ability, clinicians and educators working with this population cannot determine whether poor language skills are a result of an underlying language learning impairment or low English proficiency. From a research perspective, accurate estimation of proficiency in each language is a critical first step in every study of bilingual children. Without this, we will likely continue to find mixed results in our scientific inquiries, which are a result of the categorization of bilingual participants, rather than the question itself.

Currently, in testing for speech-language pathology, proficiency for children is determined through indirect measures (e.g., parent report of input and output of each language) with unknown or inconsistent validity (Bedore et al., 2012). Other measures that are used include rating scales, standardized testing, and language sample analysis. Though each of these measures provide some information regarding specific aspects of a child's languages, they do come with some limitations. Each of these will be addressed in the sections below.

Input and output calculations. Despite the existence of current measures to quantify proficiency and language ability, they do not provide a representation of children's true language abilities in *their functional day-to-day lives.* The most commonly used, "gold-standard" measure of proficiency in the field is a calculation of input/output in each language. Specifically, parents report exposure and use of English and Spanish (or other languages) on an hour-by-hour basis for a typical week, and average for an estimated percentage of time the child hears or uses each language.

The reason for collecting this information has been grounded in the literature. Language input has been found to be important for vocabulary development, while output has been reported to play a more critical role in syntactic development (Bedore et al, 2012) and vocabulary (Ribot, Hoff, & Burridge, 2018). There are numerous studies that support that increased exposure to and/or use of a language predict improved language outcomes in that language (e.g., Govindarajan & Paradis, 2019; Hoff et al., 2012; Ribot et al 2018; Unsworth, 2016). Though one cannot argue against the importance of exposure and use of a language on language development, the simple measurement by which it is attained and interpreted is inadequate. Indeed, Paradis and Gruter (2014) argue that the idea that input and experience "is a multi-layered construct comprised, of not only basic frequency of exposure, but also interactional qualitative factors often conditioned by familial variables" (p.12).

An input and output percentage does not give us any idea of a child's functional skills with his or her languages. We do not know whether a child can combine sentences, has appropriate vocabulary usage, or can use different tenses adequately. A child may be reported to use 70% English and 30% Spanish, yet not be able to put together a meaningful sentence in Spanish. Another child with the same percentages might be able to communicate effectively in both, with a grammatical error here and there. A third child might not understand simple directions in Spanish. The point is, for a field of professionals who do work in understanding language development and diagnosing language disorders, we should be getting more from our data collection that are reflective of the information that we are interested in – not mere percentages with arbitrary significance for what it means.

Take the following example: A child transfers to a new school from a neighboring school district. The assistant principal reports to the teacher that the child is "bilingual." What are the

possibilities in the profiles of a "bilingual" child and what the teacher can expect? See Table 1 for examples of potential children for whom this label could refer to.

Table 1: Typical cases and proficiency profiles in English language learners in English-only classrooms

	 VICKI	 CARLOS	 LUIS	 CARMEN
English Language	Poor receptive Poor expressive	Poor receptive Poor expressive	Strong receptive Strong expressive	Strong receptive Strong expressive
Spanish Language	Poor receptive Poor expressive	Strong receptive Strong expressive	Strong receptive Poor expressive	Strong receptive Strong expressive
Classroom Behaviors	Does not communicate well with others in the classroom, does poorly on schoolwork. Parents have not expressed concerns about language.	Does not communicate well with others in the classroom, does poorly on schoolwork, parents do not have any concerns with language	Can communicate with the teacher and peers fluently. Parents are concerned about language development in the home language.	Can communicate with the teacher and peers fluently. No parental concerns.

When assumptions are made about what it means to be bilingual, there are ramifications not only in the educational setting, but also in the personal lives of those individuals. For example, if the assumption is that bilinguals can speak and understand both English and Spanish, the teacher may assume that Carlos is like Carmen, in terms of their linguistic abilities. These assumptions may lead to disappointment in Carlos and may misinterpret his classroom performance as being due to intellectual or behavioral issues, rather than to a lack of strong English language skills. At face value, Vicki and Carlos would look the same in the classroom. However, Carlos could get home and communicate easily and be understood by family; he simply needs to build his proficiency in English, but otherwise has intact language skills. Vicki would struggle in all environments, and likely has a language impairment. On the other hand, Luis and Carmen will look alike in the classroom, and without insight into Luis's other language,

the parents' concerns about his heritage language might be misinterpreted by the teacher as a call for special education evaluation.

For clinicians and educators, misinterpretation of poor English skills in a bilingual child can have serious consequences. Professionals may be reluctant to diagnose a language disorder in a bilingual child because of the issue of relative language proficiency mimicking impairment. In fact, Morgan et al. (2015) show that language minority children are less likely than monolingual English-speaking peers to be identified with learning disabilities or speech-language impairments. However, strong language skills in at least one language effectively rules out the presence of a language disorder.

Language sample analysis. Language sample analysis (LSA) is one of the primary measures for assessing oral language skills in Communication Sciences and Disorders (Heilmann et al., 2008). The nature of LSA removes traditional assessment biases, making it more culturally appropriately to use for eliciting narrative samples in both English and Spanish for ELL and bilingual children (Fiestas & Peña, 2004). Children are told a story from a wordless picture book and then are asked to retell the story. This can be a challenging task for some children with lower proficiency as it requires children to "process long stretches of discourse presented auditorily, drawing inferences, building a mental model or schema and to reproduce the story using specific vocabulary, connectives, and syntactic subordination to establish coherence relationships" (Gutierrez-Clellen, 2002, pg. 180). Thus, though language sample analysis is a strong measure of narrative and language abilities, it is a task that requires a more specific and refined skill-set and is most used for assessment purposes, rather than for the purpose of determining language proficiency. This led to development of the next measure to be discussed, which uses LSA in order to specifically determine language proficiency levels.

Spanish-English Language Proficiency Scale (SELPS). The issue of quantifying language proficiency has been a persisting problem in the field of Communication Sciences and Disorders. The SELPS (Smyk, Restrepo, Gorin, & Gray, 2013) is a criterion-referenced rating scale developed in order to tackle the very problem of quantifying proficiency with Spanish-English bilinguals. This measure was developed in the form of a rating scale to be used in conjunction with a child's language sample of a story retell task. The rating scale measured syntactic complexity, lexical diversity, grammatical accuracy, and verbal fluency on a scale of 1-4 for syntactic complexity and 1-5 for the remaining measures. Though the scales demonstrated adequate reliability and validity measures, the use of the scale comes with limitations. The use of the scales required a significant amount of training for raters to reach reliable ratings before being able to independently use the measure. Additionally, the scale is only intended for use for English language proficiency, for sequential bilingual children aged 4-8.

Standardized tests in Speech-Language Pathology. A persisting challenge is that appropriate and non-biased assessments are difficult to come by for ELL or bilingual students who are suspected to have speech and/or language disorders (Caesar & Kohler, 2007). It is imperative that students be assessed for their language development in both languages, so as not to inappropriately diagnose them as having a disorder. However, there is a shortage of valid, norm-referenced assessments in languages other than English (Peña, Iglesias, & Lidz, 2001). There are a handful of standardized tests available that could be used to assess communication in each of two languages. However, many of these have little or no data to support evidence-based practice based on test development procedures. For example, many of the available tests do not report the psychometric properties of the test to demonstrate appropriate sensitivity and specificity ranges. Sensitivity refers to how accurately a test can detect a disorder, while

specificity refers to the accuracy in the test identifying those without disorders. Values above 80% for both of these measures should be the minimal amount (Plante & Vance, 1994). If psychometric properties are reported, they often do not meet these guidelines.

On the other hand, other tests with adequate psychometrics that have recently become available have a limited age range for the given population. While a strong step in the right direction for having tests designed for ELL and bilingual language speakers, there are still limitations in the students for whom the test is appropriate for. Another issue complicating the use of standardized testing is that a clinician has to assume language dominance and proficiency before administering a test, to later determine skills in the stronger language. Therefore, one would have to administer versions for both languages of the assessment in order to make a proficiency judgment. It has also been argued that norm-referenced test results may reflect how well children can take a test, rather than whether their ability to produce language is impaired (Peña & Iglesias, 1992), leading to a poor estimation of skill levels in both languages.

None of the available standardized measures address functional skills in communication. The available published measures often provide no information concerning whether a child is able to use both of his or her languages appropriately to express ideas in daily communication. This is the heart of what matters for bilingual speakers. Indeed, Grosjean (1992), argued that, "The bilingual's communicative competence cannot be evaluated correctly through only one language; it must be studied instead through the bilingual's total language repertoire as it is used in his or her everyday life" (p.472). Thus, examining children's ability to use each of their languages in a task that is familiar to them, such as translation, could provide a more accurate estimation of linguistic skills.

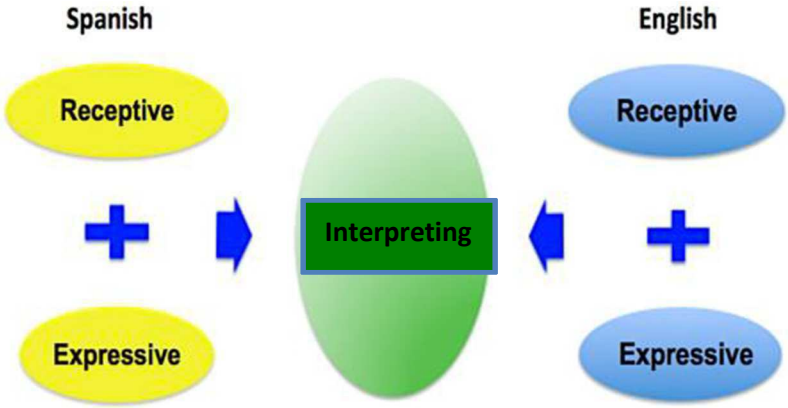
Novel approach using Theory of Natural Translation

Why might translation be a solution to the proficiency problem? Direct observation of children's ability to translate between their languages could provide a much-needed initial look at relative language proficiency in bilingual children. It has been established that clinicians need to know a child's Spanish and English proficiency in order to accurately interpret their assessments. For example, Barragan, Castilla-Earls, Restrepo, and Gray (2018) explicitly state that language proficiency will affect a child's performance on the CELF (a standardized language test) that is often used for clinical decision-making in speech-language pathology.

The starting point is the theory of natural translation (Harris & Sherwood, 1978), which refers to translation done in everyday circumstances by those who have had no special training. Many children who come from language minority backgrounds share something in common: the ability to use both languages to communicate. Children who are born to immigrant parents or who have Spanish-speaking caretakers (e.g., grandparents), often take on a role as a *language broker* (Weisskirch & Alva, 2002). As language brokers, children take on a variety of translating and interpreting tasks that provide their caretakers with crucial linkages to community information resources and other situations of daily living. The translation literature makes a distinction between the terms used to describe a written and verbal translation. When information is verbally transferred from one language to another, the term used is interpreting. In other areas, Translation with a capital "T" can be used to refer to both translation and interpreting. For the purposes of this work, we will refer to verbal translations as interpreting, from here and on.

Interpreting requires an individual to process linguistic information in one language and use the knowledge of the second language to construct a meaningful message in that language. Thus, interpreting requires an individual to incorporate all components of language (e.g., vocabulary, word order, meaning, sentence structure, pragmatics) in understanding the original message (i.e., receptive language) and incorporate all those same components to produce a similar message in the second language (i.e., expressive language). See **Figure 1**. Thus, a task using interpreting procedures follows evidence-based process of testing both languages, simultaneously, in both expressive and receptive language domains in order to convey a message, adding an important pragmatic demand. By capitalizing on this process, we can gain a better understanding of children's relative strengths and weaknesses in each of their languages.

Figure 1: Requirements for interpreting



What is important about this theory is that it predicts that a child need not have formal practice with, or training in interpreting to be able to interpret. The theory of natural translation points to interpreting as a default skill for anyone who is bilingual. Even a child in a bilingual home who does not need to interpret for anyone else will engage in 'mental-interpreting' by nature of being bilingual. For example, if a child gets home from school and their Spanish-speaking mother asks them about a book the child is reading for homework, the child would have to use what they have read about that book in English, to explain it to her mother in Spanish

(e.g., The story is about a girl who saves the world 🕮 El cuento se trata de una niña que salva el mundo). Another example of this could be a child that went on vacation in Mexico to visit family, where all language input and output is in Spanish. When the child returns to school and is asked to write a story about what she did over break, the child would have to mentally-interpret the information of their experiences that were all based in Spanish, to English (e.g., Fuimos al mar con mis abuelos y comimos mucho 🕮 We went to the beach with my grandparents and ate a lot of food). This clearly distinguishes translation of everyday activities from the job of an interpreter, in which a person must often learn specialized vocabulary to communicate more complex, specialized linguistic information. That said, for most bilingual children, translation and interpretation is part of their everyday life, with between 90 to 97% of children reported to translate and interpret for others (Tse, 1995; Weisskirch & Alva, 2002).

Despite the existing knowledge of translation and interpreting abilities in bilingual children in disciplines including psychology, translation studies, sociolinguistics, education, and cognitive science (Malakoff, 1992; Mcquillan & Tse, 2009; Morales & Hanson, 2005; Theiry, 1978; Tse, 1995), the knowledge has not translated to the development of more accurate language assessment in this population. The purpose of this work is to enhance the gold-standard in the field by creating a task to measure linguistic skills that is familiar to this population and that can yield valid representations of functional language proficiency across the bilingual spectrum. By combining the framework of natural translation (Harris & Sherwood, 1978) and a functional-componential approach to translation[1] assessment (Colina, 2008) to characterize language skills, we will be able to quantify areas of language weakness and strength in both

[1] Though we have selected the term for verbal translation as interpreting, some references to "translation" continue in this section, and subsequent sections, as formal names of theories and approaches.

languages. Functional-componential evaluation approach calls for translations to be evaluated relative to the function, or purpose of the text, and that the characteristics of the audience for whom the translated information is for to be highlighted as part of the evaluation process (Colina, 2008). The theory-driven approach to task development will also be accompanied by rigorous procedures from Test theory (i.e., Item Response Theory) to evaluate task components and total task reliability and validity.

Functional Language Proficiency (FLP) task development

The development of the task was split several stages. The first stage was the conceptual development stage, which focused on determining which presentation of the task would elicit the best response from young children ranging in age from five to nine years old. Funding and time limitations restricted the delivery tools and software that we were able to use to create a finished product. Once Powerpoint was selected as the most accessible and suitable delivery mechanism for the task, the next step was to consider what the task would look like. Considering that young children would be seeing these videos, some ideas included: 1) find cartoon characters online and do voice-overs, 2) find other young children to film that "need help", 3) find existing clips online. Though each of these options had their own merits, they were not deemed to be naturalistic, nor consistent, in terms of the feedback the viewer was receiving, and could be complicated by other information in the videos. Thus, I wanted to maintain the essence of whena child would likeliest be to perform a task like this, in daily living situations.

Orellana, Dorner, and Pulido (2003) highlight the settings in which a child might be expected to play the role of a language broker - that is, to translate from Spanish to English or vice versa for another person. These settings included: education, medical/health, commercial, cultural/entertainment, legal/state, financial/employment, and housing/residential. Orellana et al.

(2003) provided an in-depth analysis of the situations and people for whom bilingual or ELL children are translating and interpreting. However, their research was based on fifth and sixth graders, so I adjusted scenarios accordingly for the younger population for this study (i.e., at the store, at school, at the doctor's office, ordering food, asking for an appointment, and at a birthday party), expecting, for example, that younger children would not be able to have the vocabulary to assist with legal or employment issues.

Once naturalistic, age-appropriate scenarios were selected, I was mindful in making decisions about the nature of each item. For example, I took into account both lexical complexity and length of the item to translate. These two concepts are key to consider, especially for younger children. Factors like word choice (e.g., frequent v. infrequent vocabulary terms) and syntactic complexity can change the difficulty of two functionally equivalent utterances. For example, if you were to say, "he showed me how to paint the flower" v. "he demonstrated how to paint the flower", most adults could easily understand both sentences and it would have no effect on communication. However, for a child, these types of differences could affect the message's difficulty in a way that obscures the meaning, both in understanding and in generating the word. In the example given above, the child might understand the utterance with 'showed', but not the one with 'demonstrated' because of the increased linguistic complexity of the message. This is likely exacerbated for young English language learners. So, I was mindful to not create items that had specialized vocabulary (e.g., science/mathematics terms) or complex, syntactic structures that would create more difficulty with processing.

Similarly, another factor that could affect interpreting is the length of the items to be translated. For example, cognitive processes like working memory may impact the amount of information that is translated, or interpreted. Specifically, longer or more complex utterances that

tax working memory are likely to be susceptible to having key elements omitted. This is particularly true for sentences that have embedded clauses and those with multiple sentences per item. As Cowan (2000) points out, "interpreting involves the translation of spoken language in "real time", which appears to require types of attention-sharing and overloading of working memory that people generally find very difficult as conditions of information processing" (p.117). I specifically controlled for the working memory demands of each item so as not to interfere with the quality of the translation children would be producing. I limited the amount of words presented per item and manipulated some items to range in length from shortest (3 words) to longest (16 words). Because the Functional Language Proficiency task requires children to alternate from both English to Spanish and Spanish to English, I also wanted to assure that one direction of interpreting was not inadvertently made more difficult by the length of an item, leading us to incorrectly interpret results as a proficiency issue. I performed a two-tailed t-test to show that length for the items for the two conditions (Spanish to English; English to Spanish) was not different, $t(29) = 1.22$, $p = 0.22$).

Figure 2: Orientation of Characters and Scenarios

Once the items were selected, individuals were recruited to volunteer to film one-to-two-sentence videos of them saying one of the 31 items. Thirteen different individuals volunteered to film for the video. The principal investigator specifically selected what items the volunteer would say and recorded volunteers at least three times per item using an iPhone X camera. All Spanish-speaking individuals selected for the recordings were native Spanish speakers. All English-speaking individuals were native English speakers. The principal investigator selected the best video for each item, and each was organized into Powerpoint. See Figure 2 for an example of what the orientation slide looked like and Figure 3 for what it looked like during the conversational exchange. Here, the person in the middle says something that needs to be translated to Spanish (e.g., "Hi, how long are you going to be in Tucson for?"). The person on the left remains on the screen as a reminder to the child that their message needs to be in Spanish (e.g., "¿Hola, por cuánto vas a estar en Tucson?). This also assisted with the principal investigator prompting, if the child had difficulty going back and forth between languages.

Figure 3: Communication Exchange During Task

The assessment of both languages, as in an interpreting task, is considered best practice in bilingual language assessment (Bedore & Peña, 2008), but is more efficient than assessing each language separately using test

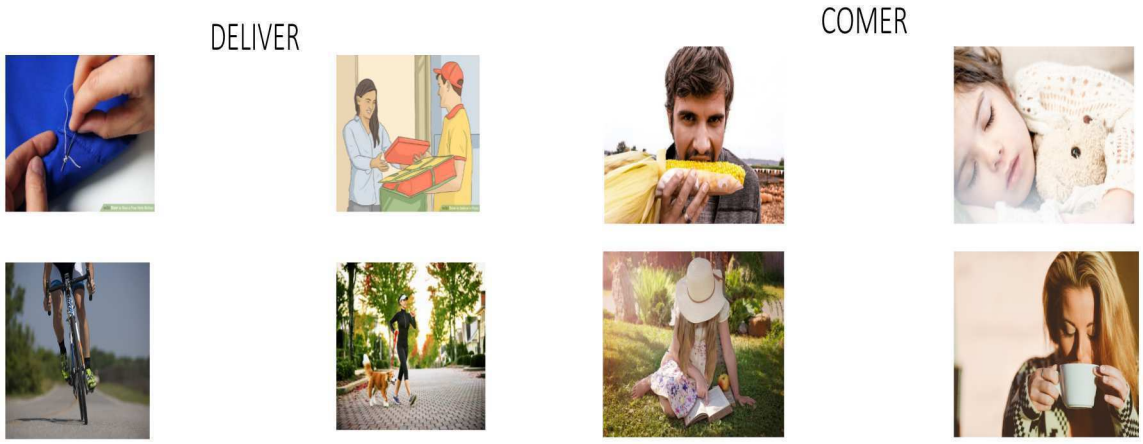
procedures. Additionally, an interpreting task holds high face and external validity, as it is generally familiar to this population in daily living.

Receptive task. In order to better interpret children's productions, we also chose to test children on task relevant vocabulary in each language to determine whether poor interpretations were related to failure to understand key vocabulary in either L1 or L2. Thus, the combination of the receptive task plus the interpreting tasks will be able to test both functional language and isolate difficulties in one language or the other.

A brief vocabulary quiz was developed for children to take after they completed the Functional Language Proficiency task. This quiz was designed to determine whether children had difficulty with their expressive language, receptive language, or both. Vocabulary quiz items were selected based on the scenarios in the task and a key word in the message that could be used. There were nine vocabulary items, with most items selected being Spanish words (2 English, 7 Spanish). When selecting the vocabulary items, we analyzed each item for potential vocabulary words to assess. However, many of the English item scenarios did not have adequate vocabulary words for testing English. Often, these potential words were abstract or not concrete. For example, we can take the item "What are you going to do when you're here?" Though there are key informational items, there is no one word that would be meaningful to isolate on its own for a measure of vocabulary. This is also true for "How long are you going to be in Tucson for?" Additionally, if items were identified with a target vocabulary word like "**soda**" in the item "Do you want to buy a large soda for two dollars more?", soda would be the same word in Spanish. These inherent differences in the items led to inequality of the items in both English and Spanish. In designing the vocabulary test, we made use of the touch-screen computer and created the task on a Powerpoint presentation. We chose to design the task for children to touch the appropriate

response, out of a field of four. Items often had semantically related items (e.g., eat vs. drink) and phonologically related items (e.g., rojo vs. ojo) in order to more closely assess children's knowledge of the item. See Figure 4 for an example an item on the vocabulary test.

Figure 4: Examples of Vocabulary Test Items

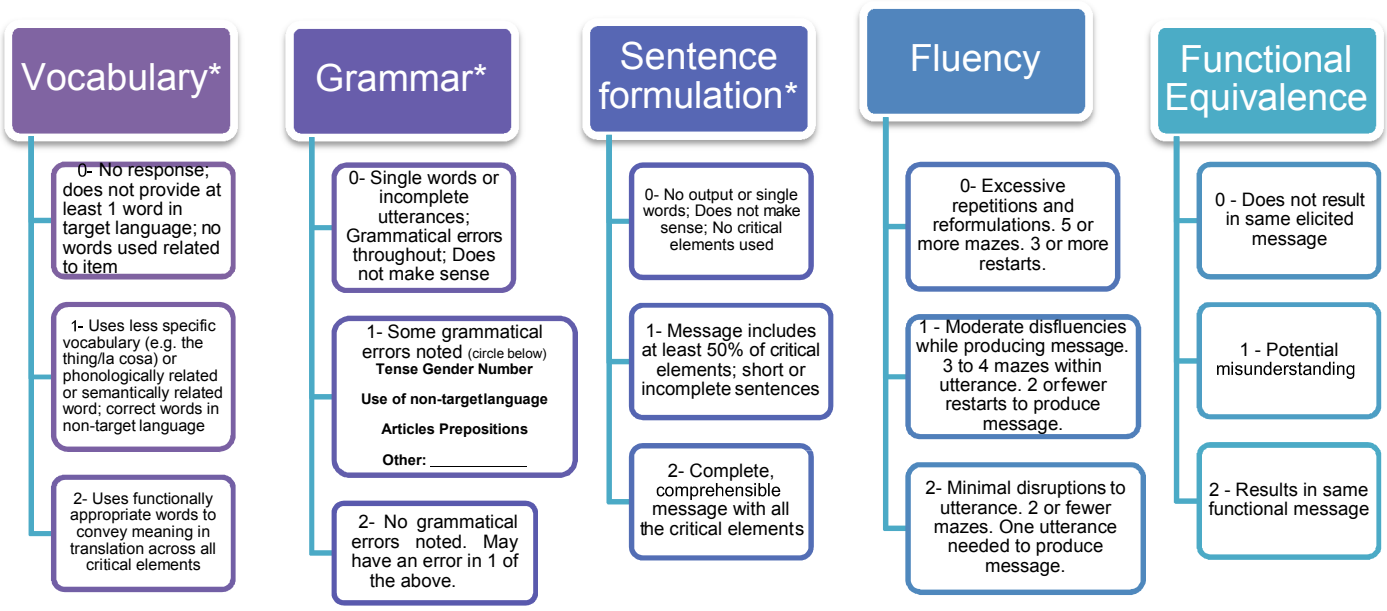


Task scoring. A scoring scheme was designed to quantify areas of language weakness and strength in both languages. This approach for evaluation of translation quality includes: evaluation that integrates different aspects of translation quality, user-defined areas of importance, and opportunity for users to select descriptive statements for classifying translation quality. The Functional Language Proficiency Task Scoring Scheme was developed by Arizmendi and Alt to follow the principles of functional translation theory. These principles state that the function of the target message is the most important criterion for translation decisions (Nord, 1991). We also take into consideration the linguistic detail needed by speech-language pathologists to accurately understand language development. See Figure 5 for the Functional Language Proficiency Task Scoring Scheme.

The ultimate goal of a functionalist approach to translation is to achieve functional equivalence (Colina, 2015). **Functional equivalence** refers to whether the message is eliciting

the same response, or function, when it is translated. For example, if I tell a child to find out "where the brooms are" and she interprets by simply saying "escobas (brooms)?", this form of equivalence may still be satisfied, even though linguistically the child did not include all the words (i.e., "*Where are the* brooms?").

Figure 5: Functional Language Proficiency Task Scoring Scheme



The context would make the 'where are the' part of the message clear without being said. Functional equivalence was rated on a 2-point rating scale to determine whether children were conveying the appropriate message to listener, despite linguistic differences noted in the translation. Satisfying functional equivalence can vary, depending on the age and experience of the communication partners. **Linguistic equivalence** was included to ensure that words and linguistic meaning used in translation, or interpreting, are the same for both the translated and original message (Peña, 2007). For example, if I say, "What time is it?", the interpreted version in Spanish would need to include a message that consists of the critical elements: 1) asking 2) time. Critical elements refer to the amount key details presented in a message, as noted in the

example above. We would know that an interpretation is not linguistically equivalent if an interpreted version asked for crayons, or took two sentences to get the same idea across. **Linguistic equivalence** was rated using a two-point rating scale on four parameters that assesses linguistic areas including: vocabulary, grammatical use, sentence formulation skills, and fluency, which should serve as a window to the child's linguistic formulation abilities. In Figure 5 above, the asterisk * next to Vocabulary, Grammar, and Sentence Formulation denotes that category's inclusion in Linguistic equivalence.

Measurement approaches to assess the reliability and validity of the FLP Task

As noted above, no measure is worthwhile, however well-intended, if it does not have appropriate psychometric qualities, reliability, or validity. The purpose of developing the Functional Language Proficiency task is to identify the child's functional proficiency, while also identifying areas of language strengths and weakness across languages in Spanish-English speaking children. By developing a reliable and valid tool for measuring language proficiency, we can be more confident in our judgments of language proficiency and have more information to understand a child's language abilities, than we currently do with input and output percentages. The theoretical and statistical approaches that I will use to develop my Language Proficiency Scale will be discussed in the following sections.

Item Response Theory and Classical Test Theory. Item Response Theory, or IRT, is comprised of statistical approaches and modeling often used in test development and in scaling abilities relative to an individual's performance on a given measure that targets a specific construct/variable. In fact, Cai et al. (2016) refer to it as being one of the "central methodological pillars supporting many large high-profile assessments around the globe (p.2)." For example, let us think about most

formal testing that we may encounter, from tests in an academic setting to those we encounter in daily living (e.g., written drivers exam). Most tests are developed so that there is a continuum of "easy" questions to "hard/er" questions. Given this, you would expect that the test takers with the highest ability, or knowledge, on the construct being tested would have the highest probability of getting both easy and hard questions correct. Under this assumption, one would also expect that those with the lowest ability, or knowledge, would get some of the easy answers correct and, conversely, have a lower probability of answering the harder questions correctly. Given a range of knowledge or ability that is being measured, we see that ability reflected in the performance on each individual test item, which eventually adds up to the final performance score. Harder questions would be weighted differently than easier questions, which is what ultimately leads to graded levels of ability on any given construct being tested. These are the general principles of IRT (Embretson & Reise, 2000).

Following these principles, there are several methods that are used to evaluate items along what typical performance would look like for those least proficient in the tested construct to those most able, and all those in between. By modeling the relationship between individual responses to specific items and ability level on the construct being tested, in this case, proficiency in a given language, the result is referred to as an item characteristic curve, or ICC. ICCs demonstrate the relationship between an individual's ability and the likelihood of them correctly answering a specific item (Hambleton, Swaminathan, & Rogers, 1991).

While IRT is one of the most used commonly used ways to measure a skill or ability, there is another form of measurement that came before IRT, called Classical Test Theory (CTT). Classical Test Theory is a theory-based measurement, based on the idea that the scores that one achieves can be broken down into true score and error. In other words, CTT runs on the premise

that the observed score on any given measure is equal to true score + error. Compared to IRT, which focuses more on the item-level information in a test, CTT focuses on test-level information (Fan, 1998). In contrast to IRT, CTT does not model a test-taker's ability to succeed on each specific item, but rather considers a pool of test-takers' performance to examine success rate on an item. Because of this, a major limitation of CTT is that the observed score is item dependent, while the item difficulty is examinee dependent (Fan, 1998). That is, item difficulty will continuously vary dependent on the ability of the pool of test-takers at any given time, though the scores generated cannot be compared between each group population. So, at any given time that the test is administered, the true score that test-takers receive can only apply to that set of test-takers. However, we also know that not all test items are created equal, which further complicates assigning a dichotomous (0-1, right/wrong) score for an "easy" item versus a "hard" item, when in the end, it equals the same amount. By doing so, gauging accurate estimates of ability becomes much more difficult.

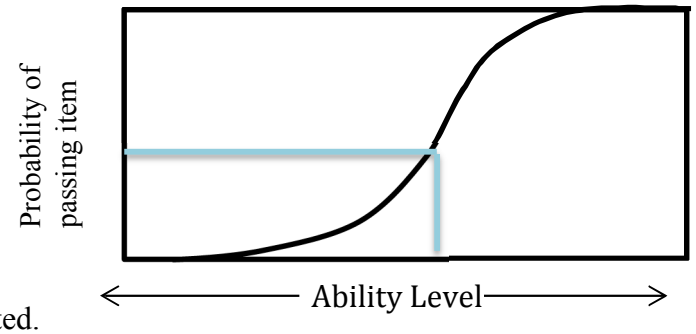
In order to create a quality assessment that captures the range of abilities or capacity that is being measured, we turn to using IRT to assess and quantify performance of Spanish- and English-speaking children on the Functional Language Proficiency task. As Cai and colleagues (2016) state, "IRT helps address technical issues such as item analysis, score reliability, and scale alignment that are related to the inherent fairness, quality, and validity questions associated with the development, administration, maintenance, interpretation, and use of tests. (p.2)" By using IRT, we have the opportunity to compare Spanish-English bilingual children, an extremely heterogeneous group, on linguistic skills in each of their languages and analyze the data based on the performance on the task compared to their peers. By having a range of proficiencyor

linguistic skills, we bypass the limitations that come with CTT, described above, and provide a more accurate picture of how to measure language proficiency skills.

Item Analysis. The basic approach will use the Rasch (1960) model to analyze the relationship between children's ability to interpret short, verbally presented information and the child's score on a 0-2 point scale for each item that was interpreted. The simple Rasch model, or one parameter logistic model (1PL), emphasizes dichotomous item formats where the dependent variable is the dichotomous item response (e.g., pass/fail), and the independent variables are the test-takers' ability score and the difficulty of the item (Embretson & Reise, 2000). Though the simple Rasch 1PL model is the most well-known model Rasch developed, he also created more complex models with increasing parameters that can be assessed. However, for our work, we use the polytomous model because ability is assessed in a 0-1-2 rating format for item difficulty across five domains. These domains are: vocabulary, grammar, sentence formulation, fluency, and functional equivalence of children's performance on the translation task. We will evaluate infit and outfit statistics and item difficulty statistics using Winsteps, a Rasch measurement software (Linacre, 2015).

Infit & outfit statistics. Additional information that can be generated from the ICC's are infit and outfit statistics. According to Linacre (2002), fit statistics are done to test specific hypotheses. Specifically, they examine how well the model and the data fit. Infit statistics refer to data that are inlier sensitive or information-weighted. This means that infit statistics are more dependent on individual patterns of responses to items

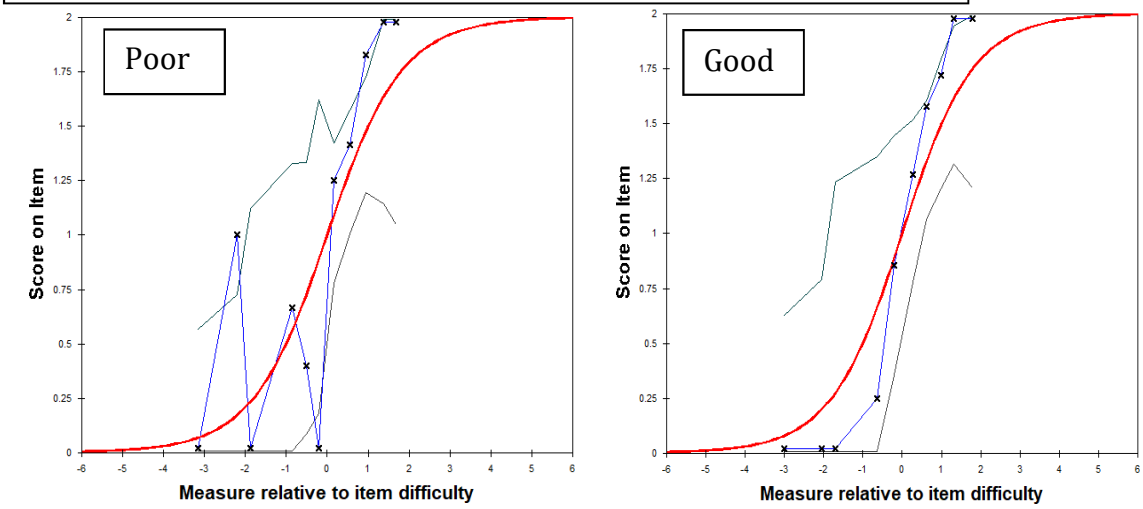
Figure 6: Rasch Model ICC Item Response



targeted on the individual and vice versa (Linacre, 2002). In other words, infit statistics can be used to assess how close an individual's ability is to the difficulty level of a specific test item. For example, a test-taker with average ability should have infit statistics on those items that other test-takers of similar ability perform at. (See Figure 6 for example of Rasch Model item response ICC curve). In this example, a person with average ability on a trait would have a 50% chance of correctly guessing this item, as it is an item with average difficulty. The higher the person's ability, the greater the likelihood of them passing that item.

On the other hand, outfit statistics refer to data that are outlier sensitive. These are more sensitive to items that show difficulty far away from a person's ability. Outfit statistics more clearly show information on where test-takers from the lower range of ability to the highest ability fit on the item characteristic curve (Linacre, 2002). For example, lucky guesses could demonstrate an underfit for the individual whose ability level is lower than the targeted item. In this case, we would see the standard ICC curve fall completely out of line with the standard curve for that test-taker (e.g., either falling completely under or over the estimated curve. See Figure 7 below for the comparison of poor infit and outfit ICC curves.

Figure 7. Examples of Poor and Good ICC curves



Generally, infit and outfit statistics evaluate how well performance on any given item fits the model on the ICC curves. These are important when we consider the nature of test development and item selection. Not all items that are generated will be thought to be "good" items, in a sense that they may not be measuring what one intended them to measure for any number of reasons. For example, a question may be poorly written and lead to a wrong interpretation of what the question was asking. In this case, most test responders may not answer the question accurately, not due to difficulty per se. An additional example would be if an item is "easy", but a significant number of high-performing test-takers unexpectedly get the answer wrong (e.g., perhaps they were overthinking it, unintended interpretation).

In order to evaluate and identify these items that do not appear to measure what we expect, we review the "fit" of each item. Though there is a range of infit and outfit from 0 to infinity with an expectation of 1 as a good fit, one guideline suggests that values from 0.5 to 1.5 are acceptable (de Ayala, 2009). However, several researchers recommend that one must take sample size into account order to accurately interpret results (Smith, Schumaker, & Bush, 1998). For this task, we will hold infit and outfit statistics to be acceptable at a cut-off of 1.5, as the purpose of the test is not diagnostic in nature, for which one would apply more conservative cut-offs. Rather, the purpose of this measure is to determine relative levels of proficiency in each language.

Upper and lower asymptotes. Asymptote measures provide more information about whether the items are measuring abilities in a sound manner. Items would ideally have lower asymptotes of 0.0, with upper asymptotes of 2.0. Note that the upper asymptote can change, depending on the measurement scale you use. For the Functional Language Proficiency task, we have a scoring rubric of 0-1-2. So, 2.0 is the highest number for our upper asymptote. For other

researchers whose rating is on a 0 -1 scale, they would seek upper asymptotes at 1.0 (Tucci, Plante, Vance, & Oglivie, 2019). A lower asymptote of 0 would mean that a low ability test taker will have zero chances of getting that item correct. The higher the lower asymptote is, the more likely the low ability test taker will get that item correct. The opposite is true for the upper asymptotes. If upper asymptotes are at 2.00 for the task, this would mean that all test-takers with the highest ability would correctly pass that item with 100% chances. For the purposes of this task, items that fell above .20 for the lower asymptote and below 1.90 for upper asymptotes will be identified, but not excluded, for determining retention in next phase of test development. The purpose for this is that the difficulty of items can be modified for content differences (e.g., the speaker in the video could have said words clearer) or for differences relating to difficulty (e.g., if the people with the highest ability who were tested cannot get the item correct, determine if more participants with even higher abilities could correctly respond).

Inter-rater reliability. Inter-Rater Reliability (IRR) refers to the relative consistency in ratings given by two or more raters of a measure with multiple items (Lebreton & Senter, 2008). The purpose of establishing IRR is to determine if raters judge items consistently with other raters. Without reliability, the validity and generalizability of the results can be misinterpreted or flawed (Tinsley & Weiss, 2000). The rating form developed for the Functional Language Proficiency task is categorical in nature, assessing five categories (Vocabulary, Grammar, Sentence Formulation, Fluency, and Functional Equivalence) on a 0-2 scale. Thus, inter-rater reliability will be quantified using a point-to-point reliability on 30% of participants.

Test retest reliability. Test retest reliability is another key factor that must be examined when developing a new measure. For the novel task, a way we examine validity is through test-retest-reliability. Test-retest reliability refers to multiple administrations of a measure to the same

people to examine the consistency of the measure itself, in order to show correlation between test scores on a test or measure. If the scores are consistent over time, it would hold true that the test is measuring what one is expecting it to measure, providing reliability and construct validity for its use.

For example, if we were to administer a test measuring nonverbal intelligence to a child in January, we would expect that if that same test were to be administered in March, the child would achieve similar scores. We could not expect a drastic change in performance (e.g., 15 point difference) in such a short amount of time. Though not likely due to error, if a test-taker obtained *exactly* the same score on the first administration and the retest, the test reliability coefficient would have perfect reliability of 1.00 (Carmines & Zeller, 1979). However, given that there are a range of factors that could lead to some variability between administrations, there is a range of reliability between .7 to 1.00 that would be considered "acceptable" to "perfect" (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). A Pearson Product Correlation will be used to obtain these results.

Language proficiency is not static. We know that depending on experience and use, it may change over time. For example, if in that time frame, a child was on summer vacation and they went to Mexico to visit family during that time, we can expect a shift in language ability. However, these are on a case-by-case basis. For the most part, children were seen during the school year. We set a period of three months in order to determine the stability of the measure after a longer period of time. For this work, we re-administered the translation task to 27 participants, comprising a third of total participants. Out of the 27 participants in 1st, 2nd, and 3rd grade – 15 participants were randomly selected from each grade for test-retest analysis.

Face validity. Face validity is a test of internal validity, and refers to whether a test appears to measure what it claims to. When a purpose of the test seems clear, by looking at it, it has high face validity. However, the inverse is true for when the purpose of a test cannot be identified – having low face validity (Nevo, 1985). Because this is a novel task, some could argue that children cannot do the task. To show face validity, we also integrated a brief four question interview at the end of the task, to ask children if they had ever had to help others understand things in different languages, and if so – for who and where. This provides some insight into child perceptions of the task and whether it was functionally relevant to their everyday experiences.

Convergent validity. Convergent validity is a test of construct validity. It assesses the degree to which a new test correlates to existing measures that test the same construct. The reasoning behind this is that two or more measures that are said to test the same thing should covary highly if they are valid measures of the concept (Campbell & Fiske; 1959). Convergent validity will be tested by comparison to language sample analysis measures, as well as input and output calculations. For the LSA measures, we will perform correlations with MLUW and Performance on the overall Functional Language Proficiency (FLP) task per language (MLUW x FLP-English, MLUW x FLP-Spanish) and predict to see a positive correlation. Additionally, we will perform correlations with Percent of Errors and Performance on the overall FLP task, per language (% Errors English x FLP-English, % Errors Spanish x FLP-Spanish) and predict to see a negative correlation.

Research questions

The purpose of this work is to develop a valid and reliable measure using interpreting as a means to uncover and quantify language proficiency in school-aged Spanish-English speakers.

The purpose of this work was to develop a novel task, grounded in theoretical and applied principles, rather than answer scientific inquiries based on specific hypotheses. However, the following questions will be addressed through this work:

1) Can children from 1st to 3rd grade perform an interpreting task?
2) Will the Functional Language Proficiency task yield strong internal validity for the measurement of language proficiency?
3) Is there any convergent validity of the Functional Language Proficiency task with existing measures, such as input and output calculations and language sample analysis?
4) Will the Functional Language Proficiency task identify a range of language proficiency abilities within children?
5) Will numerical classifications yielded from task demonstrate face validity?

Methods

Participants and recruitment. Ninety-five school-aged children participated in the study, with 32 in 1st grade, 32 in 2nd grade, and 31 in 3rd grade. Participant ages ranged from 5;11 to 9;11. Forty-eight participants were boys and 47 were girls. One hundred percent of the parents reported that their children were of Hispanic/Latino descent. Of the participants, 57 started learning English and Spanish simultaneously, that is from birth to age three, and 23 participants were early sequential learners, learning English when they entered school. There were no children who learned Spanish as a second language.

Parents and/or guardians read and signed an IRB-approved parent permission form, and provided demographic information and language development information. All parents reported

that their child knew English and Spanish. All enrolled participants were reported to be typically-developing and not enrolled in special education services. Five children were excluded from the analyses due to: parent report of language disorder/enrollment in special education (3), recent traumatic brain injury (1), or non-native speaker of Spanish (1). All children signed an IRB-approved assent form in their language of choice.

Participants were recruited by various means. The principal investigator obtained school district approval from a local school district with a high percentage of ELL and Spanish-speaking children and families. Principals were directly contacted to explain the purpose of the study, the proposed data collection procedures, and to request permission to work with children on-site after school. If a principal agreed to have his/her school participate, the principal investigator coordinated visits to each elementary school to explain the study to the children and to distribute permission form packets to the teachers to hand out and collect. All recruitment materials were written in both in English and in Spanish. Interested parents filled out consent forms, along with contact information for the principal investigator to contact directly for scheduling a visit outside of school hours. The principal investigator recruited at six different schools with principals' approval. Flyers in both English and Spanish were also posted in local libraries, community centers, and via the web.

Data collection procedures. All data collection sessions were scheduled outside of school hours. For most participants, parents chose for their child to participate in the study at their home school. Sessions were coordinated so that parents picked up students at school an hour later than their

usual dismissal time. Other options that parents could opt to choose for data collections were to participate: 1) at home, 2) at a local library, or 3) at the University of Arizona.

Data collection sessions ranged from 45 minutes to one hour. Refer to Table 2 for an outline of tasks and time to completion. The principal investigator, who is Latina and a native speaker of both Spanish and English from the Southern Arizona community, led all data collection sessions. Upon attaining the child's assent to participate, the principal investigator conducted a brief hearing and vision acuity screening. The hearing screening measured hearing of pure-tones at 500Hz, 1000Hz, 2000Hz, and 4000Hz at 20dB. To pass the vision acuity screening, children had to achieve an acuity measure of 20/40 or better (20/32. 20/25, 20/20, 20/16). Three children did not pass the hearing screening at 20 dB, but passed at 25 or 30 dB. Often, classrooms had ambient background noise that disrupted sound of the pure-tones. One child could not tolerate the hearing screening headphones on her head. These children were retained in the study, as volume was adjusted, as needed for the task to be audible.

Table 2: Data collection tasks and time to completion

1. Assent (2 minutes)	2. Vision and Hearing screening (10 minutes)	3. First language sample in language of child's choice (10 minutes)	4. Translation Task (15 minutes)	5. Vocabulary quiz (3 minutes)	6. Second language sample in other language (10 minutes)

Participants then performed a story retell task, which was used as a language sample. Children were given the opportunity to choose to do either the English or Spanish retell first. Following the Systematic Analysis of Language Transcripts (SALT) elicitation protocol developed by Miller and Iglesias (2012), the principal investigator read a script for a wordless picture book called, "Frog, Where Are You?" by Mercer Mayer (1969) in the language of the student's choice. After the principal investigator finished reading the story, the student was instructed to retell the story in that same language. Students were told that their story would be

audio-recorded and for them to use their loud, "outside" voice for better quality audio. Audio-recordings were collected using headphones with a built-in microphone connected to a Lenovo computer, using the software "Sound Recorder." Files were saved and stored for later language sample analyses described in the following sections below.

After the first language sample was collected, participants completed the Functional Language Proficiency Task. The task was presented on a Lenovo computer through a Powerpoint presentation containing 31 (total) video-clips in English and Spanish. Participants were audio-recorded once the task began, using the same headphone/microphone set that was used for the previous task. Children were instructed that they would be watching videos of two people who needed help understanding each other. Some only spoke English, while others only spoke Spanish. The participants' job was to say the message that they heard in the correct language, so that the person needing help could understand the message. For example, a child might see a video of an English-speaking person in a store saying "Hi, is there anything I could help you find?" The child would then be expected to generate a sentence that was similar in Spanish, so that the Spanish-speaking individual on the screen could understand (e.g., "¿Hola, necesitas ayuda para encontrar algo?", "¿Ocupas ayuda?", "¿Te puedo ayudar a encontrar algo?"). Once this message was relayed, the Spanish speaking person would say something in Spanish, and the child was expected to generate an utterance for the English-speaking person to understand. In this example, the Spanish speaking individual would say "¿Les puedes preguntar si tienen esta camiseta en color rojo? (i.e., "Can you ask them if they have this shirt in the color red?") Thus, each scenario required the child to facilitate a conversation between an English-only and a Spanish-only dyad. All videos showed the characters in everyday situations, including: the store, at school, at the doctor's office, ordering food, asking for an appointment, and at a birthday

party. More details on the protocol of the translation task are outlined below, under Translation Task Protocol.

Functional Language Proficiency task protocol. In the first clip, participants watch an introduction video highlighting the purpose of the task. In this video, the principal investigator says, "My name is Genesis and I'm like you. I speak English, pero también hablo Español. I mostly speak English at school y hablo español en mi casa con mi familia, but sometimes I use both to talk to my friends. What you're going to be doing today is watching videos of some of my friends who need help understanding each other. Some only speak English y otros sólo hablan español. So, the cool thing about the fact that we know both languages is that we could help them understand each other." After this, the video stopped, and the principal investigator further explained by using an example like, "When I have something I need to tell my teacher, like *I can't find my pencil* I would have to say it in English, because my teacher only speaks English. But, if I had to say that same thing to my nana, I would have to say *no puedo encontrar mi lápiz*, because she only speaks Spanish." Participants often smiled and nodded in agreement to demonstrate understanding and, at times, independently provided examples of people in their own lives who they do this for.

Participants were then oriented to the videos. At the beginning of each scenario, new characters were introduced. The person on the left hand side of the screen was always the Spanish speaker, while the person on the right was always the English speaker in order to minimize memory load of wondering which language each speaker used. The participant was told that the principal investigator would play the video clip two times to start, but that they were allowed to ask to listen to any of the clips again, as needed. After playing the clip twice, children were expected to produce the message in the other language. Children were prompted with "¿qué

dijo?" or "what did she say?" depending on the language that was expected to be used. If a child repeated the message in the same language that was originally used, the prompt expanded to, "Sí, pero ella no habla inglés. Le tenemos que decir en español. ¿Qué dijo? Ella dijo…." or "Yes, but she doesn't speak Spanish. We have to tell her in English. What did she say? She said…" If, on a subsequent prompt the child again repeated in the same language, the prompt would change to, "¿Cómo se dice en Español?" or "How do we say it in English?" The majority of the participants were able to understand after the first or second prompt, with first graders generally needing more prompting at the beginning of the task. This sequence continued for the remaining scenarios.

Once children completed the task, they were asked the following questions: 1) If the task was easy or hard, 2) Why it was easy or hard, 3) If they have ever had to help people understand each other before, 4) If yes, did they have an example of when they have helped and 5) If they liked helping. If children responded "no" or "I don't know" to question #3, the question was broken down to "You've never had to help people who only speak English understand things en Español?" or "¿Nunca le has tenido que ayudar a alguien que nomas habla Español entender cosas en English?" The younger children (i.e., 1st grade) were those who needed the question broken down more often, as it sometimes appeared that they thought the question referred to whether they had ever helped people understand via video before. Responses were audio-recorded. For the majority of the participants, the questions were asked in English. However, for those children who clearly struggled with English, or notably preferred speaking Spanish with the PI, Spanish was used. These questions were asked to get an idea of the child's thoughts about the task and whether this task was functionally relevant to their everyday experiences. These data will be reported in the Results section.

After completing all seven scenarios on the Functional Language Proficiency task, children completed a brief vocabulary quiz presented on the touch-screen computer via Powerpoint. Children saw one slide with four different images in each corner of the slide. They were instructed to point to, or touch, the correct item. For example, for "deliver", they saw an image of someone sewing, someone riding a bike, someone walking a dog, and someone delivering a pizza (Refer to Figure 4). The principal investigator would read aloud the word to identify, and the child would touch their choice on the screen. For each selection, the principal investigator tracked responses on a separate tracking sheet indicating correct responses with + or incorrect responses with -. Tracking responses was always kept out of the child's view.

Children were then asked to perform the second story retell in the other language, in order to generate a language sample. The principal investigator re-read the story in the corresponding language, following SALT protocol and elicited the child's language sample following the same procedures detailed above. Once completed, children were awarded their certificate of participation and chose their prizes. Prizes consisted of stickers, small figurinesand toys, and a snack.

Parents or guardians were met in the front office, debriefed on the project, and the principal investigator answered any questions the parent or guardian had. At this time, the principal investigator collected additional information about the child's linguistic background, using a version of The Language Experience and Proficiency Questionnaire (LEAP –Q by Marian, Blumenfeld, & Kaushanskaya (2007) that was adapted by the principal investigator for this population (See https://bilingualism.northwestern.edu/leapq/, Spanish Child Pencil and Paper version). Because this is a lengthy seven-page document to fill out, parents were originally given the opportunity to take it home, complete it, and either return it to a designated location in

the elementary school's front office or have their child drop it off at that location. However, because of poor return and compliance rate, procedures for data collection changed. Parents were advised that when they picked up their child, they would need to plan for five to 10 minutes to complete the questionnaire. Some of the questions on the LEAP-Q, as well as the formatting, were particularly confusing for some of the parents – despite the form being provided in the language of their choice. Additionally, two parents reported that they did not have the reading ability to complete the form. Here, procedures were modified yet again, with the principal investigator verbally asking these questions directly and filling out the form for the parents. The main information that was of importance to collect for this work was the age of acquisition of the child's languages, the percent input and output of each language that they spoke and heard, and parent language and education background.

Of the children who participated, 30% (27 children total) of the 90 children were randomly selected to participate in a 2nd session for test-retest-reliability. Parents were contacted three months after initial session for scheduling a 10-15 minute session at school or in the home. At this time, children only completed the translation task. Once completed, parents were debriefed on the session and the principal investigator confirmed that it would be the final session that their child would participate in for this study.

After the data collections were complete, all data were de-identified and saved on a password protected computer. All files were saved under subject numbers for processing. Data processing consisted of several different levels of processing for both the story retell tasks and the translation task. To begin, I will discuss procedures for the language samples, as there is a set system of processing set in place for running through SALT.

Data processing for language sample analysis (LSA)

The purpose of collecting language samples was to confirm normal language for the participants, as well as to provide more data for validity of the translation task. All subject numbers were entered into an excel database which documented the following steps: 1) Initial Gross Transcription, 2) Breaking the transcription into C-Units, and 3) Marking errors for SALT. Each of these stages had a double-score and a revision, of which transcribers had to reach at least 90% agreement to move the sample along to the next stage. Each child had both an English and Spanish language sample, for a total of 180 language samples needing to be transcribed. All language samples in both English and Spanish completed Step 1. However, just over 30% (33.33, 10 samples) of the 30 samples in each grade were randomly selected for establishing validity of the task (30 total).

Language samples were analyzed using the SALT 2012 Research software (Miller & Iglesias, 2012). Samples were age matched +/- 6 months. They were matched with "Total Number of Utterances" in the "Analysis Set." They were compared with the Bilingual Spanish or English Narrative Story Retell Database for the "Frog, Where Are You?" (FWAY) options. The following measures were collected and entered into an Excel sheet for later comparison: Total number of Utterances, Mean Length of Utterance in Words (MLUW), Number of Different Words (NDW), Percent of Utterances with Errors in the Sample (% Errors), Percent of Mazes in the analysis set (% Mazes). For the purposes of this initial phase of test development, the measures that were compared for validity were percent of errors (e.g., grammaticality of language sample) and MLUW for English and Spanish along with the overall performance for children on the translation task by language. Bedore, Peña, Gillam, and Ho (2010) reported that MLUW is the most widely used in clinical and research settings, as data suggests that it assists

with differentiating children with low language skills in both English and Spanish, while grammatically is reported to be positively associated with judgments of language proficiency and ability (p. 500). The remaining measures that were not compared with the overall task (NDW, % Mazes) will be used for more detailed analyses in future work.

Development of FLP task data processing procedures and methodology

As the Functional Language Proficiency task is a novel measure, the procedures and methodology went through several iterations and were continuously refined during the development process. For the task, audio-files for all participants had to be transcribed for later analysis and coding. The principal investigator transcribed all 90 files for the task. Over 20% (23.33%, 7 children per grade) of the 30 children in each grade were randomly selected for transcription reliability. A native Spanish-English speaking undergraduate research assistant completed training on double-scoring procedures at the word-by-word level. The mean word-by-word reliability was 98.54 with a range of 97.11-100%.

Once reliability was established, we moved forward to deciding which of the child's productions was going to be the one that would be scored. Often, children provided multiple attempts for how they could translate the information or would self-correct. The principal investigator and a native Spanish-English speaking undergraduate research assistant conferred and came to consensus about which item to move forward, creating a set of guiding principles. The first step was to make a decision as to what information would be counted as a maze in the utterance, following procedures from SALT's protocol. Mazes refer to filled pauses, false starts, repetitions, reformulations, and interjections. This information would be placed into parentheses ( ) to indicate that it was a maze. So, for a child who said, "She said my h* my head is hurting

mucho a lot" it would be mazed to show, "She said (my h*) my head is hurting (mucho) a lot." This procedure was followed for all utterances for every participant.

If the example above was the only thing the child produced, then that utterance would be moved forward for rating. However, if a child made several attempts, additional decisions were made. For example, for a child who said, "I'm going have time with my family – I'm going have time—I'm going to spend time with my family in the store" there are several attempts to choose from. For this production, the target message would have been something along the lines of "I am going to spend time with my family and go to the stores/go shopping." Generally, the utterances that had the most number of critical elements of the message captured was the one that was selected. In this case, the critical elements of being with family and going to a store were both said in the final attempt. Additionally, it was the most grammatical. Though the child did not produce the intended message, this was the closest to the target. All final items that were moved forward were selected and agreed upon by both the principal investigator and undergraduate research assistant. Once all item decisions were made for all 90 participants, each utterance for the task was rated using the Functional Language Proficiency Task Scoring Scheme, outlined in Figure 5 above.

It was during this time that the principal investigator and the undergraduate research assistant began making decisions as to what parts of the item counted as critical elements. As mentioned in the sections before, critical elements are the important details included in the message. The critical elements were selected in terms of the purpose of the message, or what was essential to convey the correct meaning in the interpretation. For example, for an item like ¿**Cuánto** va tardar (**para ver)** al **doctor**? (i.e., How long is it going to take to see the doctor?), the elements of "how long (cuanto)", "to see (para ver)" "doctor" are the most important

elements for conveying the message. Some of these have two words together as an element, as they are necessary in conjunction for adequately conveying the message. Other items, on the other hand, had far fewer elements needed, when compared to the overall words used in the item. For example, in response to a question asking what time the bus leaves, the speaker says, “The bus leaves at 1:05.” Here, the critical elements of the message are simply the time for the purpose of responding to the question, that is (**1):(05)**. Hence, the decision-making process for selecting critical elements was guided by the purpose of the item and the necessary details needed to adequately convey the message to the listener. For all critical elements, there was a consensus agreement between the principal investigator and the undergraduate research assistant in order to achieve consistency in rating items. See Appendix A for critical elements for each item in the task.

The principal investigator then created a rating sheet for the task, with the task items listed and a 0-2 rating scale for each category listed above. See Appendix B. The bottom of the rating sheet allowed for the rater to tally the type of grammatical errors that the child was making for each language, in order to get a sense of where the child was having the most difficulty with in each language. See Figure 8 for an example of grammar tallies. The principal investigator rated all 90 participants. Over 20% (23.33%, 21 total) of the 90 samples were randomly selected for reliability ratings by a native Spanish-English bilingual undergraduate researchassistant. Point to point inter-rater reliability was calculated at 94.95%.

Figure 8: Rating Sheet Accounting for Grammatical Errors by Language

Bueno, y a que hora salen de la escuela.	0	1	2	0	1	2	0	1	2	0
The kids are out at one.	0	1	2	0	1	2	0	1	2	0
Y a que hora se va el cameon?	0	1	2	0	1	2	0	1	2	0
The bus leaves at 1:05.	0	1	2	0	1	2	0	1	2	0

Spanish Grammar Error Types:
Tense Gender Number Non-target language grammar Articles Prepositions Other:

English Grammar Error Types:
Tense Gender Number Non-target language grammar Articles Prepositions Other:

During the rating process, there were interesting cases that made for adaptation of the rubric seen in Figure 6, as well as rules for how one would rate more complex cases. For example, for the item "Llama y ponme una cita con Ana para mañana", which would need to be translated to something along the lines of, "Call and set an appointment with Ana for tomorrow," children often made mistakes that required clarification of the scoring procedures. For example, many children would say "put a doctor's appointment for tomorrow." Not only did this add a vocabulary term (doctor), but it left out the critical piece of setting the appointment *with Ana*. How would one approach the rating for vocabulary? Clearly, a critical element was missing, but was this a vocabulary problem? Our decision, for consistency, was to allow for Vocabulary to continue to be rated as a 2 because the child used appropriate vocabulary for the message he/she gave. However, the Sentence Formulation and Functional Equivalence of the utterance would be rated as a 1 without 'with Ana'. So, though vocabulary was deemed appropriate for the message, other areas would capture the loss of information in the message. This vocabulary rating is in contrast to a child who might have used an incorrect or non-specific vocabulary word, such as "Call Ana for the thing tomorrow," or "set a doctor's point tomorrow' in which the vocabulary

scale for this item would be rated as a 1. The decision making process for the Vocabulary rating was guided by whether the participants used the correct vocabulary for *what they chose to talk about* for an item.

There were several stages of troubleshooting and consensus building on several items in order to rate systematically and by rules set in place. These were discussed and agreed upon by the principal investigator and the native Spanish-English undergraduate research assistant. Refer to Appendix C for additional scoring rubric rules and common issues per item. After all items were scored, responses were entered into an excel file prepared for analysis in the Winsteps software for identifying infit and outfit statistics of the items for retention in the task. The items that were identified as needing modifications or removal from the task, based on the set cut-offs, will be discussed in the Results section.

Results

Research question #1: Can children from 1st to 3rd grade perform an interpreting task?

First and foremost, can children as young as first grade up to third grade complete a translation-based task? Yes. All 90 participants were able to complete the task. This holds true for children with the lowest amount of Spanish or English language ability, who were able to complete at least one item on the measure. Even the five children who were excluded from the analysis sample were able to perform the task. Overall, all participants completed the task in an average of 14 minutes and 12 seconds. Third grade participants completed the task in an average of 11:44 minutes, second grade participants averaged 15:16 minutes, and first grade participants averaged 15:38 minutes.

When the participants were asked if they had ever had to help others understand information in another language, 86% of the total participants reported that they had, 96.66% of 3rd graders, 80% of 2nd graders, and 79% of 1st graders). With prompting, all children reported helping for a range of people and situations. The majority helped family members (e.g., mom, dad, grandparents, siblings, aunts, uncles, cousins), followed by friends, and church members. See Table 3 for performance on the FLP task per grade

Table 3: Means and Ranges of performance on the FLP task per grade

	Percent completed	Mean FLP-English score	SD FLP-English	Range of FLP-English	Mean FLP-Spanish score	SD FLP-Spanish	Range of FLP-Spanish
1st Grade	100%	114.56	29.70	5-138	115.62	33.23	7-147
2nd Grade	100%	115.1	25.15	33-146	118.2	39.50	1-155
3rd Grade	100%	129.96	24.78	9-145	138.3	30.61	12-157

*FLP English = Spanish to English score, Maximum score –150
*FLP Spanish = English to Spanish score, Maximum score – 160

Research question #2: Will the interpreting task yield strong internal validity for the measurement of language proficiency? (Internal validity)

Item Analysis. Items were analyzed using the Winsteps software. Items were separated into English items and Spanish items in order to determine internal consistency and reliability per language direction. Chronbach's Alpha was calculated and reported, as it is a measure of internal consistency and scale reliability, with many methodologists recommending a minimum α coefficient of between 0.65 and .80 to be considered "good" (Goforth, 2015). Both English and Spanish items achieved strong consistency of items. Root Mean Square of Error (RMSE) was also calculated. Results are reported in Table 4.

Table 4: Item Analysis English and Spanish Results

	English Items	Spanish Items
Child Reliability (Chronbach's Alpha α)	.90	.92
RMSE for Child Reliability	.23	.22
Item Reliability	.88	.87
RMSE for Item Reliability	.23	.19

Analysis of infit/outfit statistics and the upper and lower asymptotes of the items was performed. Items that yielded up to 1.50 in infit and outfit statistics and lower and upper asymptotes to .50 were retained,. From the 75 items in the Spanish to English set, 56 (75%) of the items met the criteria for retention in the task, and 19 items (25%) did not meet the set cut-offs for infit and outfit statistics. See Table 5 for Spanish to English Items that were identified for modification or exclusion from next phase of the test.

Table 5: Spanish to English items with unacceptable item-level statistics to be removed or modified from future test versions.

	Infit	Outfit	Lower Asymptote	Upper Asymptote
2g. Les puedes preguntar si **tienen esta camiseta en color rojo?**	1.81			
2f. Les puedes preguntar si **tienen esta camiseta en color rojo?**	1.88	2.44		
4f. Si, pero preguntale **si me la pueden mandar a mi casa.**	1.83	3.91		
6sf. Llamale al salon y **ponme una cita con Ana para manana.**				1.41
6f. Llamale al salon y **ponme una cita con Ana para manana.**	1.78	2.11	.34	1.82
8g. Dile que **esta bien, vamos a las cinco.**	1.51			
14f. Empezó cuando me pegue con la puerta del carro.		1.69		
16fe.Voy estar aqui por tres dias.	1.62			
18f. Voy a pasar tiempo con mi familia y ir a las tiendas.		1.64		
20v. Si, por favor.	1.81		.49	
20g. Si, por favor.	2.19			
20f. Si, por favor.	1.87		.37	

20fe. Si, por favor.	2.05			
23v. Puedes pedir una pizza mediana con pepperoni y unas alitas de pollo?		2.80	.59	
23g. Puedes pedir una pizza mediana con pepperoni y unas alitas de pollo?	1.87		1.15	
23sf. Puedes pedir una pizza mediana con pepperoni y unas alitas de pollo?		2.02	.61	1.73
23f. Puedes pedir una pizza mediana con pepperoni y unas alitas de pollo?	1.78	2.71	.51	1.81
25f. No, nomas la pizza y las alitas. Tambien preguntale por los paquetitos de chile y queso.	1.73	2.11	.24	1.74
30f. ¿Y a qué hora se va el cameon?		1.62		

Of these 19 items that fell above the cut-offs set for the task, two items evaluated vocabulary, four evaluated grammar, nine evaluated fluency, two evaluated sentence formulation, and two for functional equivalence.

The same item analysis was performed for English to Spanish Items. From the total of 80 items in the English to Spanish set, 61 (77%) of the items met the criteria for retention. We identified 19 items (23%) as having unacceptable item-level statistics. See Table 6 below.

Table 6. English to Spanish items with unacceptable item-level statistics to be removed or modified from future test versions.

	Infit	Outfit	Lower Asymptote	Upper Asymptote
1v. Hi, is there anything I can help you find?		2.71	.36	
5g. Yes, we can have that delivered to your house.	1.57		.21	
5sf. Yes, we can have that delivered to your house.			.37	
5f. Yes, we can have that delivered to your house.	2.54		.35	
5fe. Yes, we can have that delivered to your house.		1.54	.43	
7g. We only have an opening for today at five.			.29	
7f. We only have an opening for today at five.	1.69	2.77	.35	1.75
9f. Hi, please fill out this form and give it to the doctor when you see him.		1.59		1.86
11f. They'll call you back in about ten minutes.	1.68	1.61		
13f. When did her head start to hurt?	1.51			
21v. We have ice cream too. Do you want				1.66

chocolate, vanilla, or both?				
24f. Do you want to buy a large soda for two dollars more?	1.73	2.14		1.84
24fe. Do you want to buy a large soda for two dollars more?				1.82
26f. It's going to be ready in ten minutes.		1.51		
27g. Can you tell this parent that we have early release tomorrow?	2.04	1.80		
27f. Can you tell this parent that we have early release tomorrow?	1.72	1.73		1.86
31v. The bus leaves at 1:05.		1.88	.43	
31f. The bus leaves at 1:05.	2.01	2.05	.38	
31fe. The bus leaves at 1:05.		3.25	.42	

Of the 19 items that did not fall within the cut-offs, three items evaluated vocabulary, three items evaluated grammar, one evaluated sentence formulation, nine evaluated fluency, and three evaluated functional equivalence.

Test retest reliability. A total of 15 samples were tested for test-retest reliability. The Spearman's rho Correlation Coefficient was used using SPSS. Test retest coefficient fell at .876. This indicates good reliability (Heise, 1969) between the first administration and the second administration of the task, three months apart.

Research question #3: Is there any convergent validity of the interpreting task with existing measures, such as language sample analysis and input and output calculations?

Language sample analysis measures. To test the validity of the task with existing measures, we performed two correlation analyses with the overall score of the participants on the task to their Mean Length of Utterance in Words (MLUW) and to their Percent of Errors in each language with performance on the task. Only 35% (32 participants) of the total participants' (90) language sample measures were compared to performance on the functional proficiency task, for both English and Spanish measures. For the English Measures of MLUW and percent of errors compared to the task, the data are presented in Table 7. Percent of errors on the English language

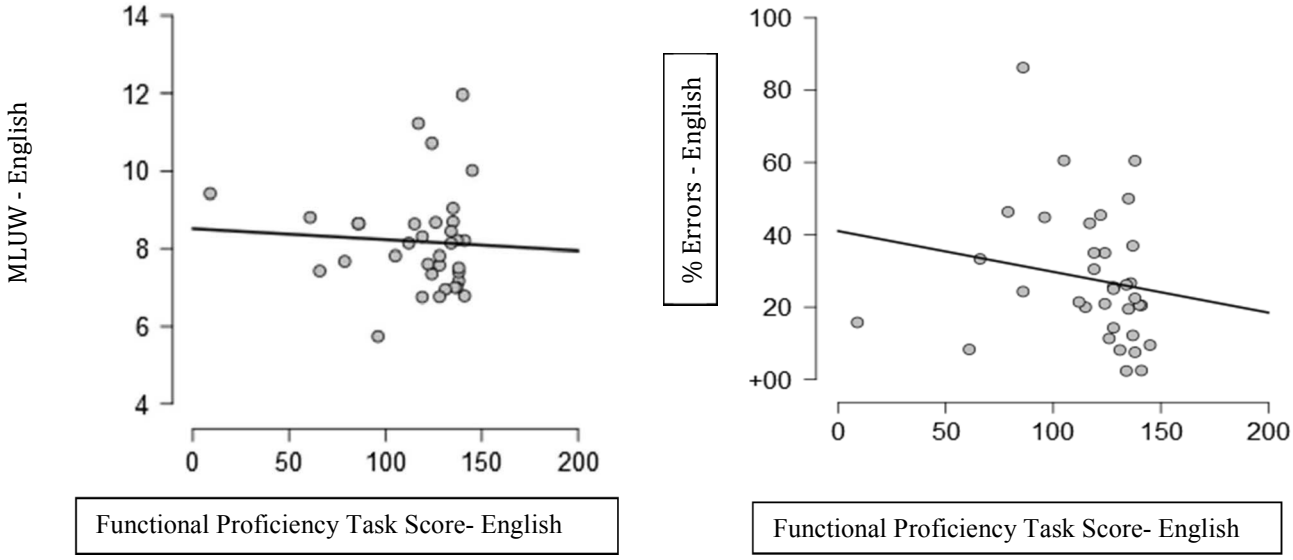
samples was significantly correlated with the Functional Proficiency Task in English. MLUW was not correlated with the Functional Proficiency Task in English.

Table 7: English Language Sample Measure correlations with Functional Proficiency Task

	Spearman's rho	p
MLUW and Functional Proficiency Task	-0.114	0.742
Percent of Errors x Functional Proficiency Task	-0.332	0.026

See Figure 9 below for distribution of correlations for MLUW and Percent of Errors in English with Functional Proficiency Task.

Figure 9: MLUW and Percent of Errors correlations with English Functional



The same correlation analyses were conducted for the Spanish language sample measures (MLUW and % Errors) with the Functional Proficiency Task. These data are presented in Table

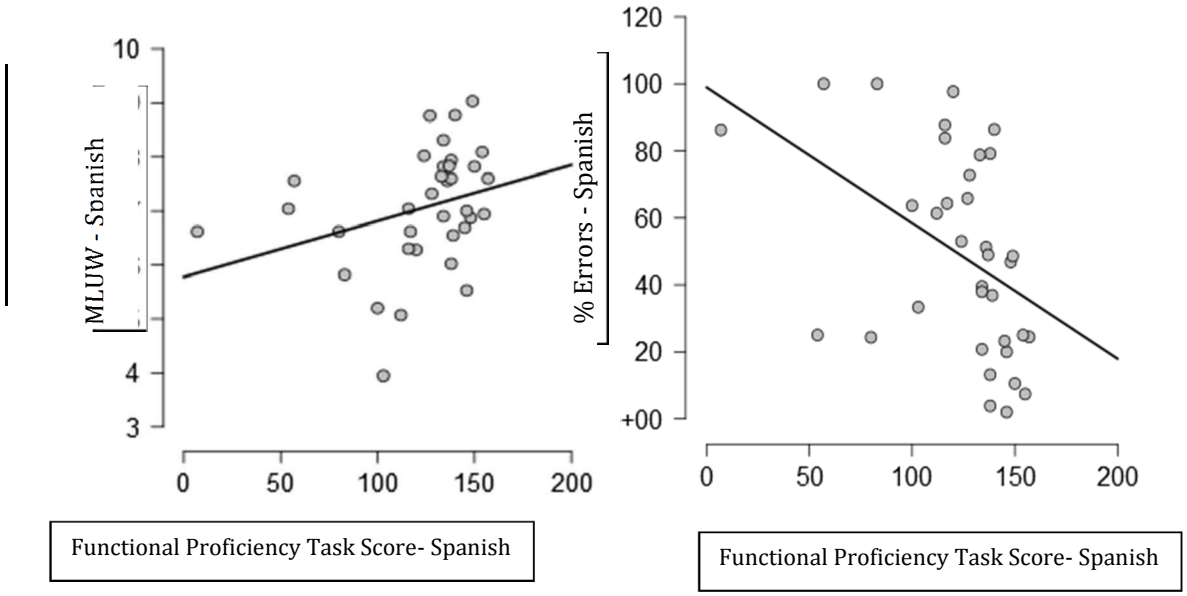
8. Both MLUW and Percent of Errors in Spanish were significantly correlated with the Functional Language Proficiency Task.

Table 8: Spanish Language Sample Measure correlations with Functional Proficiency Task

	Spearman's rho	p
MLUW and Functional Language Proficiency Task	0.387	0.011
Percent of Errors x Functional Language Proficiency Task	-0.560	<.001

See Figure 10 below for distribution of correlations for MLUW and Percent of Errors in Spanish with Functional Proficiency Task.

Figure 10: MLUW and % Error correlations with Spanish Functional Proficiency Task



Input and output calculations. Parent report of input and output was organized into an Excel sheet and performance on the Functional Language Proficiency task. When we examine the relationship between these two measures, there is some consistency with performance. However, it is often mixed for what one would predict based on input and output calculations. It is important to note that when we examine the Input compared to Output alone, we also see an

inconsistency in predictions (e.g., a child who hears 50/50 of each language but output is 90/10). There are many examples of these inconsistencies throughout the complete sample, which can be found in Appendix F. However, for reporting in this section, the top 10 children with highest English output and 10 children with the lowest English output are compared. See Table 9 below for example of children with highest English output to those with the lowest English output, and their performance on the FLP task.

Table 9: Input and Output comparison to FLP ranked from highest to lowest English Output

Subject	Input E	Input S	Output E	Output S	FLP-English	FLP-Spanish
1242	90	10	100	0	77	6
1351	95	5	100	0	9	12
1223	50	50	90	10	141	155
1312	80	20	90	10	128	133
1315	90	10	90	10	145	157
2116	70	30	85	15	134	117
1119	50	50	80	20	112	139
1124	80	20	80	20	128	120
1142	70	30	80	20	86	80
1353	50	50	80	20	136	112
2231	60	40	40	60	117	134
2264	50	50	40	60	131	141
2381	50	50	40	60	134	150
1126	50	50	30	70	81	100
1373	20	80	30	70	135	151
1355	20	80	25	75	115	116
2351	50	50	20	80	133	153
2363	50	50	20	80	140	155
2364	30	70	20	80	139	152
1225	50	50	5	95	134	138
1363	50	50	5	95	131	146

FOURTH QUARTILE Superior	THIRD QUARTILE Advanced
SECOND QUARTILE Intermediate	FIRST QUARTILE - Minimal

Here we see that Subject 1363 (the last row of the table), who is reported to speak Spanish 95% of the time, achieves, as expected, highest performance in the fourth quartile for the

FLP-Spanish. However, what would not be expected, based on reported output of 5% in English is that this child also achieved advanced proficiency measurement on the FLP-English. If we only look at input for Subject 1363 (50/50), we would expect this result and categorization on the FLP task. However, by looking at output, we would not.

For further analysis, we compared input and output, LSA measures, and performance on the FLP task. All participants are reported in Table 10 below.

Table 10: Participant Input/Output, Language Sample Analysis Measures, and FLP Performance

Subject	Input E	Input S	Output E	Output S	MLUW E	MLUW S	% Errors E	% Errors S	FPT E	FPT S
1351	95	5	100	0	9.42	NA	15.79	NA	9	12
1142	95	5	95	5	8.65	6.62	86.21	24.32	86	80
1315	90	10	90	10	10.02	7.6	9.52	24.44	145	157
1312	80	20	90	10	7.58	7.64	25.58	78.79	128	133
2116	70	30	85	15	8.14	6.62	2.38	64.29	134	117
1353	50	50	80	20	7	5.07	26.67	61.36	136	112
1124	80	20	80	20	6.76	6.28	14.29	97.67	128	120
1119	50	50	80	20	8.14	6.55	21.43	36.84	112	139
1142	70	30	80	20	8.65	6.62	24.32	86.21	86	7
2141	50	50	80	20	8.81	7.56	8.33	100	61	57
1317	50	50	70	30	8.21	6.87	20.51	46.81	141	148
1352	70	30	70	30	10.72	8.77	20.93	86.36	124	140
1141	50	50	70	30	7.35	3.95*	35	33.33	124	103
2222	50	50	70	30	7.68	NA	46.34	NA	79	69
1341	50	50	60	40	11.97	9.03	20.59	48.57	140	149
2122	70	30	60	40	8.7	7.6	50	79.17	135	138
1125	50	50	60	40	NA	7.94	NA	13.16	133	138
1112	30	70	60	40	8.68	8.76	11.36	65.79	126	127
1114	60	40	60	40	8.31	7.04	30.51	87.69	119	116
1121	40	60	55	45	7.43	5.82	33.33	100	66	83
1211	20	80	50	50	6.78	6.94	2.5	7.41	141	155
1318	50	50	50	50	7.17	7	22.45	2	138	146
2313	50	50	50	50	7.4	5.53	7.55	20	138	146
2227	20	80	50	50	7.51	7.82	60.47	10.53	138	150
2121	70	30	50	50	7	7.56	12.2	51.22	137	136
1116	50	50	50	50	8.22	6.69	36.96	23.19	137	145
1313	50	50	50	50	9.05	8.09	19.51	25	135	154
1117	50	50	50	50	6.95	7.82	8.16	39.47	131	134
2229	70	30	50	50	7.82	7.83	25	48.94	128	137
1113	50	50	50	50	5.74	5.2	44.83	63.64	96	100

1224	70	30	40	60	8.45	6.02	26.19	3.85	134	138
2113	40	60	40	60	7.61	8.02	45.45	52.94	122	124
1151	50	50	40	60	6.75	7.32	35	72.73	119	128
2231	60	40	40	60	11.23	8.31	43.18	37.93	117	134
2223	40	60	40	60	7.82	6.9	60.53	20.75	105	134
1355	20	80	25	75	8.64	6.3	20	83.78	115	116
1225	50	50	5	95	NA	7.04	NA	25	33	54

INPUT/OUTPUT and FLP COLOR CODING	
FOURTH QUARTILE - Superior	THIRD QUARTILE Advanced
SECOND QUARTILE Intermediate	FIRST QUARTILE Minimal

LSA MEASURES COLOR CODING	
2SD above mean (MLUW)	1SD within mean
1SD below mean	2SD above mean (Errors)

Note that for MLUW, being above 2SD is positive (green)
For Percent of Errors, being above 2SD is negative (orange)

These data are ranked from highest English output to lowest, for the sample of 30% of children for whom LSA measures were calculated. This table continues to highlight the discrepancies between input/output calculations and overall performance across measures. For example, Subject 1315 (3rd row from the top) was reported to both hear and speak English 90% of the time, and hear and use Spanish 10% of the time. On the LSA measures, he was 2SD above the mean for MLUW English, within 1SD for MLUW-Spanish and Percent of Errors in English, and above 2SD for Percent of Errors in Spanish. When we see performance on the FLP, this subject attains the highest level of performance for Functional Language Proficiency in both languages.

Research question #4: Will the interpreting task identify a range of language proficiency abilities within children?

To answer this research question, child performance scores were calculated by adding all scores. For English, the scores could range from 0-150, for Spanish, the scores could range from 0-160. Scores were then ordered from highest to lowest, and separated into quartiles using the Excel quartiles function. The Quartiles are reported below in Table 11.

Table 11: Quartiles for Performance on Novel Task by language

	English	Spanish
First Quartile – Minimal	117.25	117.25
Second Quartile - Intermediate	131	136
Third Quartile - Advanced	137	146
Fourth Quartile - Superior	146	157

Additionally, subject numbers were ranked from highest to lowest performing on each language, in order to determine within group/quartile performance on task. An adequate representation of range in performance was found. See Appendix D for complete list of participant performance by quartile, per language.

Research question # 5: Will numerical classifications yielded from task demonstrate face validity?

When we consider the data presented in Table 10, we see that children demonstrate a range of proficiencyfor the Functional Language Proficiency task.We selected children in each grade, in each quartile, as a qualitative comparison to demonstrate what their performance would look like on a task. Children in each quartile either fell into one of four categories: Superior, Advanced, Intermediate, and Minimal Functional Proficiency. See Tables 12, 13, 14, and 15 for Response Examples per category and quartile.

Table 12: Fourth Quartile Response Examples –Superior Functional Proficiency

1st Grade	2nd Grade	3rd Grade
"I'm gonna spend time with my family and go to the stores."	"She's gonna spend time with her family and go to the store."	"I'm gonna spend time with my family and go shopping."
"Tenemos nieve. ¿Quieres chocolate, vainilla, o las dos?	"Tambien tenemos nieve. ¿Quieres (um) chocolate, vanilla, o los dos?"	"Tenemos nieve tambien. ¿Quieres vainilla, o chocolate, o quieres los dos?"

Table 13: Third Quartile Response Examples – Advanced Functional Proficiency

1st Grade	2nd Grade	3rd Grade
"She is gonna stay with her family and go to the shop."	"I'm going spend time with my family at the store."	"She going to have time with her family and in the store."
¿Quieres nieve? Tenemos	"Tiene nieve, si quiere vanilla o	"Tiene ice cream tambien y dijo

chocolate y vanilla (quieres los mismos) ¿Quieres los dos?	chocolate y dos?	que y si quiere vanilla o chocolate o los dos?"

Table 14: Second Quartile Response Examples – Intermediate Functional Proficiency

1st Grade	2nd Grade	3rd Grade
"I'm gonna start time with my family and I'm gonna go to the stores."	"(im gonna start) I'm gonna start with my families going (to the) to the stores."	"(I'm gonna) I'm gonna have time with my family and go (to) to stores.
"(Quieres uh uh) ¿Quieres nieve? (tamb*) Tengo nieve también. ¿Quieres nieve, vanilla o fresa o y ya?"	"¿Quieres (um) ice cream de (vani*) vanilla y chocolate?	"Tenemos (um) nieve tambien. (quie*) ¿Quieres vanilla chocolate o la dos?

Table 15: First Quartile Response Examples—Minimal Functional Proficiency

1st Grade	2nd Grade	3rd Grade
"I going to go with (con) mi family and i go (to the) to the X tiendas."	"I will go to the store with my family."	"I will planned a birthday party for my family."
"I know how to say ice and cream."	"Chocolate y nieve."	"Que el ice cream el vanilla chocolate o (l*) both?"

Discussion

The purpose of this dissertation was to develop a novel, valid, and functional task to measure language proficiency in Spanish-English speaking school-age children. By bridging the knowledge across disciplines, we developed a task that had 1) high face validity, 2) high construct validity, 3) high internal consistency, and 4) convergent validity with language sample analysis. There has been a long-standing problem for decades on finding operational definition for bilingualism, including how researchers, clinicians, and educators quantify different levels of proficiency. The focus of this dissertation was to create a task that was culturally appropriate and familiar to the population, and to determine if using such measure could give us any insights into a English-Spanish speaking child's linguistic development in each language. It is important to

highlight that this is the first phase in the development of the task. However, this work yielded positive results and outcomes from which to continue building upon in future work.

Quantity and quality. It makes sense as to why input and output has long been considered a strong metric for language proficiency. The argument has been made that the more one listens to and uses a language, their language abilities in that language should also be reflective of it. When we consider the input and output calculations that were reported by parents in this study, we found examples across the board for how these estimates are not adequate. Why would it be that there were children who were reported to listen to English and Spanish an equal amount of time, yet are reported to use one language 95% of the time? Why would another who is reported to hear and use both languages equally (i.e., 50/50 input, 50/50 output) be attaining a score of minimal proficiency on the FLP? Why is it the case that children with identical input/output score have drastically diverging scores on both LSAs and the FLP task? These are only a couple examples, out of many that were found in this study, that highlight that the percentage of time language is heard is not always reflective of use and ability.

What could be contributing to this discrepancy in the input and use of languages? Not all input is quality input (Hoff & Core, 2013). We know that it is both quantity and quality of input that leads to improved language outcomes (Vigil, Hodges & Klee, 2005). Another factor that might be more meaningful for a population with divided language input and output is the proportion of time the child is listening to language versus speaking each language. Take for example, a child with the output estimates 70% English, 30% Spanish. The child might get home from school and watch TV the rest of the day, without having to *use* language to communicate with others. This is after being at school hearing *and* using English, for classroom tasks, to work with peers, during recess. Once home (all Spanish speaking), the child likely does not need to

use language to the same degree as the needs are for use at school. The child may respond to questions at a very basic level (e.g., "¿Cómo estuvo la escuela? – "Bien.", "¿Tienes hambre?- Sí.", "¿Qué quieres comer ahora? – No sé."). These basic conversations are likely common among parents and their children. It has been well-documented in the literature that young children's screen time exceeds recommendations (e.g., Atkin, Sharp, Corder, & van Slujus, 2014; Carson & Janssen, 2012; Hale & Guan, 2015) with up to an average of four hours/day for children as young as preschool (Lauricella, Wartella, & Rideout, 2015). Indeed, excess screen time is now associated with parents reporting having family meals less than four days of the week. (Gingold & Schoendorf, 2014). It's one thing to be listening to one language on television or a tablet for five hours a day, but another when you are both hearing and using them to achieve a purpose, like in conversation or as needed in an academic environment. The fact that the percent of the time that children are listening to and using language isn't yielding accurate representations in language measures may be due to the differences in proportions of language listening and language use in each language. For example, if a child hears and uses English equally at school, there might be a 50/50 ratio of hearing **and** use of English. However, if at home, the child hears more Spanish, than actually uses it, it could be a ratio of 80 (hearing)/ 20 (using). These are different representations of the child's language than the percent of time heard and used in each language, per hours of the day.

This difference in actual proportion of time hearing and using language, between languages, may also be a contributing factor in how English and Spanish seem to capture different results for what measures are significant. As we saw with the FLP task, performance on Spanish was correlated with MLUW and Percent of Errors in Spanish with LSA. For English, only Percent Errors in English was correlated with the FLP task. This has been seen time and

time again in the bilingual literature, with different measures being meaningful for one language, but not another (e.g., Alt, Arizmendi, & DiLallo, 2014; Bedore et al., 2011;). Perhaps the children in this sample who are required to actually *use* more language in Spanish (e.g., those that are interpreting more regularly for their parents) are different from those who do not need to as often. Furthermore, the proportion of time that they hear and use both languages in English, is relatively constant for the better part of the day for these children at school (i.e., which requires children to both listen to and use language at comparable rates). This may be why we may not be finding convergent validity with the English FLP and MLUW in English. These are empirical questions to test in future studies.

While input and output calculation estimates are interesting, they cannot be taken with confidence that they are accurate. It cannot be true that for every single hour reported for every single day, the child will be listening or using the language that was reported for these calculations. It is simply not realistic to say that those measures would hold true for every single day of that child's life. What if one day the child makes a Spanish-speaking friend? What if another day Spanish-speaking cousins are in town? What if that weekend all the child hears is English while having a sleepover at a friend's house? The input and output measurement that is argued to be indicative of language proficiency does not provide us with much information on language, and is not stable. This was evident when we compared input/output reports alone for these participants. Though some of the input and output percentages did match up with predicted performance on the FLP, many were far off from where one would estimate. For example, we would expect a child who is exposed to and uses both languages equally (50/50 input and output) would attain a superior or advanced FLP score. There were some instances where children with these percentages attained a minimal score on the task. Why? Because they did not have enough

language ability in one language to successfully complete the task. This is telling in and of itself as to how much bilingual language ability the child has. If they score in the minimal ranges due to very limited language skills in one of the languages, this would mean that the child is functionally monolingual. One would assume that if you listen to and use both languages equally, you would have strong bilingual language skills. This is where the input and output reports become problematic. Certainly, there were children who did match up with the parent report of input and output. However, when there are so many examples and inconsistencies between the reports and the performance on the task, we know that it is an underlying issue of poor representation of language skills through the use of input and output percentages.

Functional Language Proficiency task advantages. We discussed in earlier sections that there are other available measures that can be used for determining proficiency, each with merits and limitations. What this task brings to the table that is different from previous measures, is that it assesses functional language and how it is used in English and Spanish speaking children's day to day lives. It capitalizes on using a task that is familiar to them. This cannot be said about the majority of assessment practices. Additionally, it provides insights into what parents are hearing at home, and why perhaps there might be significant concerns in one language over another when hearing the differences in their child's language abilities (e.g., "He doesn't speak Spanish like my nephew who's the same age"). This task allowed us to identify who those children were. For example, for some children, the task was not challenging. However, it is important to note that none of the participants achieved full points on the FLP for either language. For others, it was easier in one direction over another. Some struggled with grammar, others with formulating sentences, and others struggled with overall fluency in trying to get a message across. As we

know about individuals who speak more than one language, their abilities are heterogeneous and can constantly be shifting depending on their experiences.

On the task, we identified children across the range of proficiency levels and separated them into quartiles for classification. Children were spread almost evenly across categories in the quartile ranges for both languages. The range began with children who were only being able to complete one item (i.e., Si por favor/ Yes, please) to those who could readily complete every item on the task with minimal hesitations. As expected, children did not perform equally in both languages. There were some children who attained scores in the same category (e.g., Superior English FLP and Superior Spanish FLP), but many evidenced mixed proficiency across languages. This is the nature of bilingualism.

Additionally, we see an incremental growth from 1st to 3rd grade in mean scores attained on the task. The 3rd grade children were those who had the highest average scores on the FLP, as would be expected with more experience using their languages. It was key in this stage of development to find the range of language abilities for children who are reported to be bilingual by their parents. All the children in the study were able to complete the task and was reflective of their language skills when we compared to the language sample analysis measures.

When we think about the children that were presented in Table 1, we can now identify how each child would perform on the Functional Language Proficiency Task. See Table

See Table 15 for Typical proficiency profiles with performance on the Functional Language Task. Refer to Tables 8 and 9 for comparison with measures and performance.

Table 16: Typical proficiency profiles with performance on the Functional Language Task.

	 VICKI	 CARLOS	 LUIS	 CARMEN
English Language	Poor receptive Poor expressive	Poor receptive Poor expressive	Strong receptive Strong expressive	Strong receptive Strong expressive
Spanish Language	Poor receptive Poor expressive	Strong receptive Strong expressive	Strong receptive Poor expressive	Strong receptive Strong expressive
FLP Outcome	Not tested Future directions	Minimal FLP-E Minimal FLP-S e.g., Subject 1225	Advanced FLP-E Intermediate FLP-S e.g., Subject 2115	Superior FLP-E Superior FLP-S e.g., Subject 1315

With the use of this task, we can have a better representation of who these children are and what they can functionally do with both their languages in a mere 15 minutes. It is important to note that while Carlos does have strong Spanish skills, he was not able to complete the task because he did not understand what the person was saying in English. Thus, all translations going from English to Spanish could not be interpreted. Additionally, though he understood the items that went from Spanish to English, his poor expressive English language also limited him from transmitting that message in English. Something to point out here is that the default categorization on this task would lead Carlos to have "Minimal Functional Proficiency" in Spanish, this is not true and should be taken in the context of this task. This could be something to refine or factor in for improving the task to more appropriately categorize children that would fit this profile.

Future directions. There are many areas that can be explored from the data gathered from this dissertation.These directions can range from task development and refinement to testing of specific questions that we can answer from the data that has already been collected. Recall, that this is a robust sample size of children (90 total, 30 per grade), from which we can draw more accurate conclusions about the task and its utility.

Task refinement. One area would be to improve the Functional Language Proficiency Task. The task can be refined, through modification of poor items. Some items might need modifications such as rewording. Others might benefit from the way the item was delivered in the video (e.g., if a speaker said something too fast, or too quietly). Alternatively, poor items can be discarded from future versions, in order to make the task shorter and more efficient. However, as was reported in the Results section, participants finished in 15 minutes or less.

The task can also be refined, and expanded to apply to a larger range of proficiency ability and developmental levels (e.g., 4th grade-5th grade). This would mean following the principles outlined in this work, while also taking into account the differences that might emerge in the older-elementary school years (e.g., attrition, increased academic vocabulary knowledge). However, it would be worthwhile to pilot the current measure to apply to these other groups and see what their resulting ICC curves and measures look like. It could be the case that the items do not need to be refined, and the "harder" items that children of high ability missed in this version are correctly interpreted by the older sample, and that items that are currently unacceptable would become acceptable. Alternatively, it could be that increased amount and time in school results in a loss of language, or attrition of Spanish, resulting in more errors and looking similar to the younger children in this sample. In this case, the poorly-fitting items would need to be re-worked, replaced, or discarded. These would all be worthwhile factors and questions to explore.

Functional Language Proficiency task comparisons to LSA. For the purposes of this work, we only tested MLUW and Percent of Errors to overall proficiency scores in each language. However, our LSA data also include measures of vocabulary (number of different words), fluency (percent of mazes), on top of the measures of sentence formulation (MLUW) and grammaticality (percent of errors) that were used to compare to the task *as a whole.* Because of

this, each specific LSA measure can be specifically tested for what it corresponds to on the Functional Language Proficiency Task. That is, we can test for convergent validity of vocabulary ratings in one language to the measure of Number of Different Words; sentence formulation ratings in one language to the MLUW. Perhaps by testing these language domains separately, we may find more convergence on the task as a whole to LSA measures.

Additionally, recall that correlations were only examined with the performance on the Functional Language Proficiency Task for 30% of children for whom we calculated SALT measures for. We may find more meaningful differences with the full set of 90 participants. Additionally, there may be more specific language profiles from which to start categorizing the level of functional language proficiency.

Differences in performance in children with developmental language disorders.

Another key area to explore is how Spanish-English speaking children with developmental language disorders perform on the Functional Language Proficiency Task. If performance on this measure yields multiple indicators for differences with the typically-developing group, it would mean that it would be a worthwhile measure to develop into a test not only of language proficiency, but for *diagnosing* language disorder. As has been discussed in earlier sections, there are few valid and reliable measures available in the field of Communication Sciences and Disorders, specifically for the Spanish-English speaking population. Being able to extend this task as an assessment measure would be invaluable to correct diagnoses of language disorder vs. language proficiency. We have discussed the repercussions that this could have on this field, special education, and education. It is imperative that we begin to design more valid, reliable, and functional measures in order to prepare for the increasingly diverse cultural and linguistic generations to come.

Appendix A. Critical Elements for Functional Language Proficiency Task Items

Critical Elements in **Bold,** Parentheses denotes that the combined words count as one critical element.

Item	Number of Elements
1. Hi, is there **anything** I can **help** you **find**?	3
2. ¿ **Tienen esta camiseta** en color **rojo**?	4
3. Sorry, we **only** have that one in **black**. Do you **want** us **to order** one for you?	4
4. **Si** me la **pueden mandar** a (**mi casa)**.	4
5. **Yes**, we can have that delivered to your house.	1
6. Ponme una **cita** con **Ana** para **mañana**.	3
7. We only have an **opening** for **today** at **five**	3
8. Esta bien, **vamos** a las **cinco**.	2
9. Hi, please **(fill out) this** form and **give** it to the **doctor** when you see him.	4
10. ¿**Cuánto** va tardar (**para ver)** al **doctor**?	3
11. They'll call you back in about **ten minutes**.	2
12. Me ha estado **doliendo** mucho la **cabeza**.	2
13. **When** did her head **start** to **hurt**?	3
14. Empezo cuando (**me pegue)** con la **puerta** del **carro**.	3
15. Hi, **(how long)** are you going to (**be in) Tucson** for?	3
16. Voy estar aqui por **tres dias**.	2
17. **What** are you **going (to do)** when you're **here**?	4
18. Voy a **pasar tiempo** con (**mi familia)** y ir a las **tiendas**.	4
19. Do you **want** some **cake**?	2
20. **Si**, por favor.	1
21. We have **(ice cream)**, too. Do you **want chocolate, vanilla,** or **both**?	5
22. **Nomas chocolate**. No me gusta la vainilla.	2
23. Puedes pedir una **pizza mediana** con **pepperoni** y unas **alitas** de pollo?	4
24. Do you **want** to buy a **large soda** for **two dollars** more?	5
25. **No**, nomas la pizza y las alitas. Tambien preguntale por los **paquetitos** de **chile y queso**.	5
26. It's going to be **ready** in **ten minutes**.	3
27. We **have early release tomorrow.**	4
28. Bueno, y ¿a (**qué hora) salen** de la **escuela**?	3
29. The kids are out at **one**.	1
30. ¿Y a (**qué hora)** se **va** el **cameon**?	3
31. The bus leaves at (**1):(05)**.	2

Appendix B: Functional Language Proficiency Task Scoring Sheet

FORM B: Subject # _____ **RATER:** _____

Item	Vocabulary			Grammar			Sentence Formulation			Fluency			Functional Equivalence		
1. Hi, is there anything I can help you find?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Les puedes preguntar si **tienen esta camiseta en color rojo?**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Sorry, we only have that one in black. Do you want us to order one for you?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Si, pero preguntale **si me la pueden mandar a mi casa.**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Yes, we can have that delivered to your house.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
2. Llamale al salon y **ponme una cita con Ana para manana.**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
We only have an opening for today at five.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Dile que **esta bien vamos a las cinco.**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
3. Hi, please fill out this form and give it to the doctor when you see him.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Puedes preguntarle **cuanto va tardar para ver al doctor?**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
They'll call you back in about ten minutes.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Dile que **me ha estado doliendo mucho la cabeza**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
When did her head start to hurt?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Empezo cuando me pegue con la puerta del carro.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
4. Hi, how long are you going to be in Tucson for?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Voy estar aqui por tres dias.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
What are you going to do when you're here?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Voy a pasar tiempo con mi familia y ir a las tiendas.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Do you want some cake?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Si, por favor.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
We have ice cream, too. Do you want chocolate, vanilla, or both?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Nomas chocolate. No me gusta la vainilla.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
5. Puedes pedir **una pizza mediana con pepperoni y unas alitas de pollo?**	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Do you want to buy a large soda for two dollars more?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
No, nomas la pizza y las alitas. Tambien preguntale por los paquetitos de chile y queso.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
It's going to be ready in ten minutes.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
6. Can you tell this parent that **we have early release tomorrow**?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Bueno, y a que hora salen de la escuela.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
The kids are out at one.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Y a que hora se va el cameon?	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
The bus leaves at 1:05.	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2

Spanish Grammar Error Types:
Tense Gender Number Non-target language grammar Articles Prepositions Other:

English Grammar Error Types:
Tense Gender Number Non-target language grammar Articles Prepositions Other:

Notes:

Note: The bolded words denote the part of the message that was scored. (e.g., "Tell him that","Dile que" are not the parts of the message that matter for the interpretation.

Appendix C: Guidelines for selecting item to move forward for scoring

When a child produces multiple attempts for one item, you need to decide which one will be the item that is scored.

The first step is to make a decision as to what information would be counted as a maze in the utterance. **Mazes** refer to filled pauses, false starts, repetitions, reformulations, and interjections. This information would be placed into parentheses ( ) to indicate that it was a maze. The number of mazes also helps with later decisions in scoring for fluency. These follow the principles of transcription used for the Systematic Analysis of Language Transcripts (SALT).

So, for a child who said, "She said (my h*) my head is hurting (mucho) a lot."

In general: these are the principles to follow.

1) Choose the item that had the most critical elements included in the message.
2) If they hold the same number, follow it by the one that most readily achieves functional equivalence. That is, which one is likely going to achieve the same intended meaning in the target language?
    a. Sometimes, the item will not meet functional equivalence because of the child did not produce the correct target message in any of the attempts. See Example #1 below.
3) If there are still items to pick from, choose the most grammatical.
4) If, at this point, there are still items to pick between, that have met the above criteria – pick the one that has the least amount of fluency errors.

Examples: (highlighted refers to the item that was moved forward for rating)

1. He said that if he can order um some – he said if you can order him a pepperoni pizza and some salad.
    a. *Here, the child was supposed to ask for a medium pepperoni pizza with chicken wings. The last production was chosen because it was 1) the one with the most critical elements 2) the most grammatical, and 3) had the least amount of fluency errors.*

2. How much is it -- how much time is gonna take for him – how much time its gonna take for the doctor?
    a. *Here, the child made three attempts. The child was supposed to ask how much time it would take to see the doctor, or how much longer they would need to wait. The last attempts has more critical elements present, more likely will achieve functional equivalence, has some grammatical errors, but overall the most sound choice as we go down the principles to follow.*

3. (Cuando) cuando (le le le) le duele-- cuando (le) le empezo a doler la cabeza a ella – cuando le estaba doliendo la cabeza?
    a. *Here, there child also made three attempts. In the second, all critical elements are present. It's grammatical. It has a fluency error (le) maze, but we have already satisfied the requirements for the principles to follow with this child's attempt. The reason the last one was not chosen, was because it changes the meaning of the utterances to (from when did her head start to hurt vs. when was her head hurting?)*

Appendix D: Scoring Rubric Rules - Common issues

1. If a child only repeats the item in the same language after several prompts, score vocabulary as 0, judge grammar of the utterance, sentence formulation, and fluency as is, score 0 for functional equivalence.

2. Be sure to look out at any differences between the intrepretations that could affect meaning and overall message (e.g., pizza vs. pizzas would result in 1 in functional equivalence)

3. When judging fluency, do not count the first (um) or other filler (e.g., que) that is at the **beginning** of utterance.

4. It is not necessary to judge the beginning of the utterance (e.g., he said that…..she said that…) nor the fluency reformulations before the message begins. For example, if a child said "(he he said he she said that) **her head hurts a lot.**" Another example would be "(um that que) **si quiere que le ordene uno**?" You would judge the bolded message.

5. The use of "a" in Spanish in examples like "quieren **a** comprar / quiere **a** ordenar / quiere **a** buscar" is not considered an error. Dialectal and community used among Spanish-English speaking *children* in the community.

<u>Item-specific common issues:</u>

Sí, pero preguntale si me la pueden mandar a mi casa.

- "Yes, if you can **give it to** my house." Rate Vocabulary as 1, semantically related but not correct. Functional Equivalence would still be a 2.
- "Yes, but if you can order it **at** my house." Rate Vocabulary as 2, Grammar as 2 (1 error, at), Sentence Formulation as 2, Fluency as 2, Functional Equivalence as 1 (potential misunderstanding).

Ponme una cita con Ana para mañana.

- Score Vocabulary as 2 if appointment and tomorrow are included, but missing Ana. Proceed to score Sentence Formulation and Functional Equivalence as 1 without Ana.

- If child uses meeting vs. appointment, this is a semantic difference – so Vocabulary would be judged as a 1. Without the context of "llamale al salon or llamale al doctor (as many children assumed), using meeting would be appropriate. However, the child hears the context and should use appointment accordingly. Functional Equivalence would be rated as a 2 or 1, depending on how child uses the words.

We only have an opening for today at five.

- If only missing **today** - Score vocabulary as 2, but score Sentence Formulation and Functional Equivalence as 1.

Hi, please fill out this form and give it to the doctor when you see him.

- If 'firma' or 'llena' or 'completa' are missing, Vocabulary is rated as a 1.
- Papel instead of forma or formulario would be rated as a 2 in Vocabulary.

Puedes ordenarme una pizza mediana con pepperoni y unas alitas de pollo?

- If only missing **medium**,
  Score Vocabulary, Sentence Formulation, Functional Equivalence if only **medium** is missing.

Appendix D: Scoring Rubric Rules – Common issues (continued)

Puedes ordenarme una pizza mediana con pepperoni y unas alitas de pollo?

- If says child says "chicken" but not "chicken wings" (e.g., he wants a pepperoni pizza with chicken), it could be mistaken as wanting chicken on the pizza. For the example above, Vocabulary would be 1, Grammar would be a 2, Sentence Formulation would be a 1, Fluency would be a 1, and Functional Equivalence would be a 0, as medium was also missing from there. If medium was included, it would still be a 0.

Do you want to buy a large soda for two dollars more?

- If child omits (**large) grande**, vocabulary is judged as a 1 due to missing critical error. Judge Grammar, Sentence Formulation, and Fluency accordingly. Functional Equivalence would be rated as a 1.

The bus leaves at 1:05

- Que el cameon se va a **las** una cinco. Do not count the use of 'las' as en error. Dialectal and commonly used in the community.

Appendix E: Quartile ranges and scores for all participants for English and Spanish FLP measures

Subject	FPT-S	FPT-E	Subject
1315a	157	146	1226b
2372b	157	145	1315a
2361b	156	145	2228b
1211a	155	144	2261b
2363b	155	143	1371b
1313a	154	143	2361b
2351b	153	142	1362b
2364b	152	142	2362b
2261b	151	141	1317a
2365b	151	141	1211a
1373b	151	141	2372b
2227b	150	140	1341a
2374b	150	140	2363b
1371b	150	139	2214a
2381b	150	139	2365b
2371b	150	139	2364b
1341a	149	138	2313a
1314a	148	138	1318a
1317a	148	138	2111a
1128b	147	138	2227b
2313a	146	138	2268b
1318a	146	137	2121a
2228b	146	137	1116a
2362b	146	137	1127b
1363b	146	137	2374b
1226b	146	136	1353a
2251b	146	136	1356b
2112a	145	136	1152b
1116a	145	135	1313a
1321b	145	135	2122a
2267b	144	135	1373b
2214a	141	135	2371b
1356b	141	135	2251b
2264b	141	134	2116a
1372b	140	134	2381b
1352b	140	134	1224b
1119a	139	133	1125a
2262b	139	133	2351b
2122a	138	132	2211a

1125a	138	132	2114a
1362b	138	132	1262b
1224b	138	131	1117a
2268b	138	131	1129b
2229b	137	131	2267b
2121a	136	131	1363b
2127b	136	131	2264b
2231a	134	130	2115a
2111a	134	130	2112a
1117a	134	129	1372b
2223a	134	128	1312a
2115a	134	128	1314a
1251b	134	128	1124a
1312a	133	128	2229b
1262b	133	128	2127b
2123b	130	127	2263b
2211a	128	127	2262b
1151b	128	126	1112a
1112a	127	126	1321b
2114a	125	126	2243b
1152b	125	124	1141a
2113a	124	124	1352b
2226b	124	122	2113a
1129b	123	121	2266b
1127b	122	120	1251b
1124a	120	119	1114a
2225b	119	119	1151b
2263b	118	118	1128b
2116a	117	117	2231a
2266b	117	115	1355a
1261b	117	112	1119a
1355a	116	110	2225b
1114a	116	110	2123b
2243b	114	109	2265b
1353a	112	106	1223b
1141a	103	105	2223a
1113a	100	102	2226b
1126b	100	97	2311a
1223b	95	96	1113a
2124b	85	96	1263b
1121a	83	95	2124b

1142b	80	86	1142b
2265b	77	84	1261b
2222a	69	81	1126b
2311a	57	79	2222a
2141a	57	77	1241b
1225b	54	66	1121a
1351b	12	61	2141a
1142b	7	33	1225b
1241b	6	9	1351b
1263b	1	5	1142b

Appendix F: Input and output calculations compared to Functional Language Proficiency Task

Subject	Input E	Input S	Output E	Output S	FLP E	FLP S
1242	90	10	100	0	77	6
1351	95	5	100	0	9	12
1223	50	50	90	10	141	155
1312	80	20	90	10	128	133
1315	90	10	90	10	145	157
2116	70	30	85	15	134	117
1119	50	50	80	20	112	139
1124	80	20	80	20	128	120
1142	70	30	80	20	86	80
1353	50	50	80	20	136	112
2141	50	50	80	20	61	57
2225	80	20	80	20	110	119
2228	80	20	80	20	145	146
2243	80	20	80	20	126	114
2266	80	20	80	20	121	117
1263	50	50	75	25	96	1
2127	75	25	75	25	128	136
2265	50	50	75	25	109	77
2268	75	25	75	25	138	138
1128	50	50	70	30	118	147
1141	50	50	70	30	124	103
1251	70	30	70	30	120	134
1262	50	50	70	30	132	133
1317	50	50	70	30	141	148
1352	70	30	70	30	124	140
1371	80	20	70	30	143	150
1372	70	30	70	30	129	140
2222	50	50	70	30	79	69
2263	50	50	70	30	127	118
2361	70	30	70	30	143	156
1112	30	70	60	40	126	127
1114	60	40	60	40	119	116
1125	50	50	60	40	133	138
1321	60	40	60	40	126	145
1341	50	50	60	40	140	149
2122	70	30	60	40	135	138
2124	60	40	60	40	95	85
2261	80	20	60	40	144	151
2267	50	50	60	40	131	144
1121	40	60	55	45	66	83
1113	50	50	50	50	96	100
1116	50	50	50	50	137	145
1117	50	50	50	50	131	134

1129	60	40	50	50	131	123
1152	50	50	50	50	119	128
1211	20	80	50	50	136	125
1226	50	50	50	50	33	54
1261	50	50	50	50	84	117
1313	50	50	50	50	135	154
1318	50	50	50	50	138	146
1356	50	50	50	50	136	141
1362	60	40	50	50	142	138
2121	70	30	50	50	137	136
2123	50	50	50	50	110	130
2226	50	50	50	50	102	124
2227	20	80	50	50	138	150
2229	70	30	50	50	128	137
2262	50	50	50	50	127	139
2313	50	50	50	50	138	146
2362	70	30	50	50	142	146
2365	50	50	50	50	139	151
2371	50	50	50	50	135	150
2372	60	40	50	50	141	157
2374	60	40	50	50	137	150
1151	50	50	40	60	5	7
1224	70	30	40	60	106	95
2113	40	60	40	60	122	124
2223	40	60	40	60	105	134
2231	60	40	40	60	117	134
2264	50	50	40	60	131	141
2381	50	50	40	60	134	150
1126	50	50	30	70	81	100
1373	20	80	30	70	135	151
1355	20	80	25	75	115	116
2351	50	50	20	80	133	153
2363	50	50	20	80	140	155
2364	30	70	20	80	139	152
1225	50	50	5	95	134	138
1363	50	50	5	95	131	146

FLP COLOR CODING	
FOURTH QUARTILE	THIRD QUARTILE
SECOND QUARTILE	FIRST QUARTILE

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational assessment*, *8*(3), 231-257.

Alt, M., Arizmendi, G. D., Beal, C. R., & Hurtado, J. S. (2013). The Effect of Test Translation on the Performance of Second Grade English Learners on the KeyMath □ 3. *Psychology in the Schools*, *50*(1), 27-36.

Alt, M., Arizmendi, G. D., & DiLallo, J. N. (2016). The role of socioeconomic status in the narrative story retells of school-aged English language learners. *Language, speech, and hearing services in schools*, *47*(4), 313-323.

Arizmendi, G. D., Alt, M., Gray, S., Hogan, T. P., Green, S., & Cowan, N. (2018). Do bilingual children have an executive function advantage? Results from inhibition, shifting, and updating tasks. *Language, speech, and hearing services in schools*, *49*(3), 356-378.

Artiles, A., Rueda, R., Salazar, J., & Higareda, I. (2005). Within-Group Diversity in Minority Disproportionate Representation: English Language Learners in Urban School Districts. *Exceptional Children, 71*(3), 283-300.

Atkin, A. J., Sharp, S. J., Corder, K., van Sluijs, E. M., & International Children's Accelerometry Database (ICAD) Collaborators. (2014). Prevalence and correlates of screen time in youth: an international perspective. *American Journal of Preventive Medicine*, *47*(6), 803-807.

Barragan, B., Castilla-Earls, A., Martinez-Nieto, L., Restrepo, M. A., & Gray, S. (2018). Performance of low-income dual language learners attending English-only schools on the Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish. *Language, speech, and hearing services in schools*, *49*(2), 292-305.

Bedore, L. M., Pena, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of communication disorders*, *43*(6), 498-510.

Bedore, L. M., & Pena, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, *11*(1), 1-29.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education*, *15*(4), rm4.

Caesar, L. G., & Kohler, P. D. (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools*.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. *Annual Review of Statistics and Its Application, 3*(1), 297-321.

Campbell, D., Fiske, D., & Helson, Harry. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Sage publications.

Carson, V., & Janssen, I. (2012). Associations between factors within the home setting and screen time among children aged 0-5 years: A cross-sectional study. *BMC Public Health, 12*, 539.

Cowan, N. (2000). Processing limits of selective attention and working memory. Potential Implications for interpreting. *Interpreting*, *5* (2), 117-146.

De Ayala (2009) The Theory and Practice of Item Response Theory. *Psychometrika, 75*(4), 778-779.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, *58*(3), 357-381.

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children. *Language, Speech, and Hearing Services in Schools*. *35*(2), 155-168.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363-385.

Gingold, J., Simon, A., & Schoendorf, K. (2014). Excess Screen Time in US Children: Association With Family Rules and Alternative Activities. *Clinical Pediatrics, 53*(1), 41-50.

Goforth, C. (2015, November 16). *Using and interpreting Chronbach's Alpha.* Retrieved from https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/

Govindarajan, K., & Paradis, J. (2019). Narrative abilities of bilingual children with and without Developmental Language Disorder (SLI): Differentiation and the role of age and input factors. *Journal of communication disorders*, *77*, 1-16.

Grosjean, F. (1985). The bilingual as a competent but specific speaker hearer. *Journal of Multilingual & Multicultural Development*, *6*(6), 467-477.

Hale, L., & Guan, S. (2015). Screen time and sleep among school-aged children and adolescents: a systematic literature review. *Sleep medicine reviews*, *21*, 50-58.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

Hammer, C. S., Detwiler, J. S., Detwiler, J., Blood, G. W., & Qualls, C. D. (2004). Speech–language pathologists' training and confidence in serving Spanish–English Bilingual children. *Journal of communication disorders*, *37*(2), 91-108.

Heilmann, J., Miller, J. F., Iglesias, A., Fabiano-Smith, L., Nockerts, A., & Andriacchi, K. D. (2008). Narrative transcription accuracy and reliability in two languages. *Topics in Language Disorders*, *28*(2), 178-188.

Heise, D. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review, 34*(1), 93-101.

Hoff, E., & Core, C. (2013). Input and language development in bilingually developing children. In *Seminars in speech and language, 34*(4), 215-226.

Hoff, E., & Core, C. (2015). What clinicians need to know about bilingual development. *Seminars in speech and language, 36*(2), 89-99.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of child language*, *39*(1), 1-27.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of communication disorders*, *43*(6), 456-473.

Lauricella, A. R., Wartella, E., & Rideout, V. J. (2015). Young children's screen time: The complex role of parent and child factors. *Journal of Applied Developmental Psychology*, *36*, 11-17.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, *11*(4), 815-852.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Menken, K. (2008). *English Learners Left Behind : Standardized Testing As Language Policy*. Clevedon: Multilingual Matters. Retrieved from http://ezproxy.library.arizona.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=222241&site=ehost-live

Miller, P. C., & Endo, H. (2004). Understanding and meeting the needs of ESL students. *Phi Delta Kappan*, *85*(10), 786-791.

Morgan, P., Farkas, G., Cook, M., Strassfeld, N., Hillemeier, M., Pun, W., Schussler, D. (2018). Are Hispanic, Asian, Native American, or Language-Minority Children Overrepresented in Special Education? *Exceptional Children, 84*(3), 261-279.

Miller, J. & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC."

National Center for Education Statistics (2018, April) English language learners in public schools. Retrieved from https://nces.ed.gov/programs/coe/indicator_cgf.asp

Nevo, Baruch. (1985). Face Validity Revisited. *Journal of Educational Measurement,22*(4), 287-293.

Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology*.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15-24.

Powers, S., Johnson, D. M., Slaughter, H. B., Crowder, C., & Jones, P. B. (1985). Reliability and Validity of the Language Proficiency Measure. *Educational and Psychological Measurement*, *45*(4), 959–963.

Ragan, A. & Lesaux, N. (2006). Federal, state, and district level English language learner program entry and exit requirements: Effects on the education of language minority learners. *Education Policy Analysis Archives, 14*, 20.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Ribot, K. M., Hoff, E., & Burridge, A. (2018). Language use contributes to expressive language growth: Evidence from bilingual children. *Child development*, *89*(3), 929-940.

Roseberry-McKibbin, C., Brice, A., & O'Hanlon, L. (2005). Serving English language learners in public school settings. *Language, speech, and hearing services in schools*.

Rumberger, R., & Larson, K. (1998). Toward Explaining Differences in Educational Achievement among Mexican American Language-Minority Students. *Sociology of Education,71*(1), 68-92.

Miller, J. & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Steinberg, L., & Thissen, D. (2013-04-23). Item Response Theory. In (Ed.), The Oxford Handbook of Research Strategies for Clinical Psychology. : Oxford University Press,. Retrieved 2 Aug. 2018, from

http://www.oxfordhandbooks.com.ezproxy1.library.arizona.edu/view/10.1093/oxfordhb/978 0199793549.001.0001/oxfordhb-9780199793549-e-018.

Sullivan, A. (2011). Disproportionality in Special Education Identification and Placement of English Language Learners. *Exceptional Children, 77*(3), 317-334.

Tinsley, H. E., & Weiss, D. J. (2000). Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). Academic Press.

Tucci, A., Plante, E., Vance, R., & Oglivie, T. (2019). Data-driven item selection for the Shirts and Shoes Test. *Journal of communication disorders*, *78*, 46-56.

Unsworth, S. (2016). Early child L2 acquisition: Age or input effects? Neither, or both?.*Journal of child language*, *43*(3), 608-634.

U.S. Census Bureau, Population Division. (2018, June). Annual Estimates of the Resident Population by Sex, Age, Race, and Hispanic Origin for the United States and States: April 1, 2010 to July 1, 2017.

Vigil, D. C., Hodges, J., & Klee, T. (2005). Quantity and quality of parental language input to late-talking toddlers during play. *Child Language Teaching and Therapy*, *21*(2), 107–122.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*(3-4), 139-159.